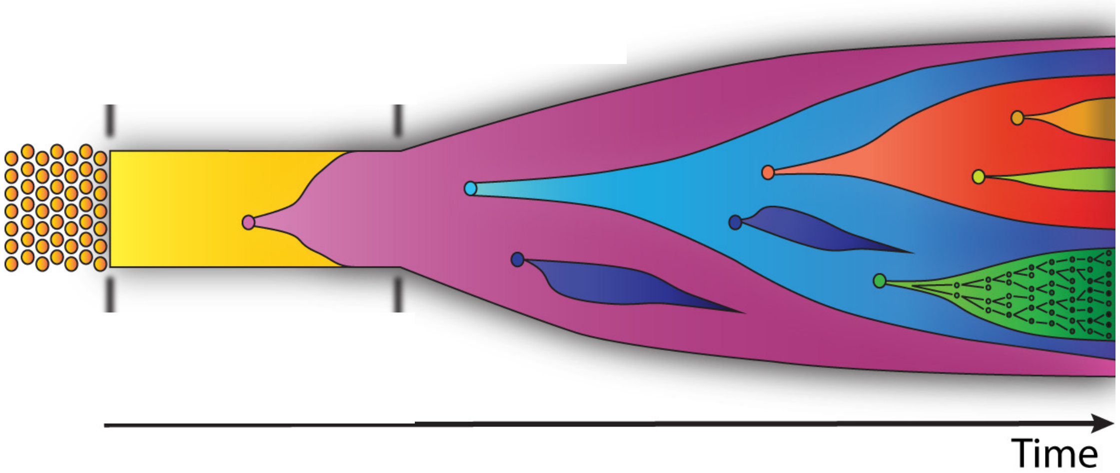
Royal Institute of Technology

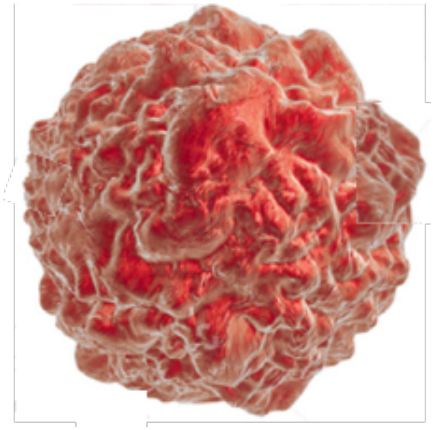
DD2434 STAT. ADVANCE MACHINE LEARNING HT 2017

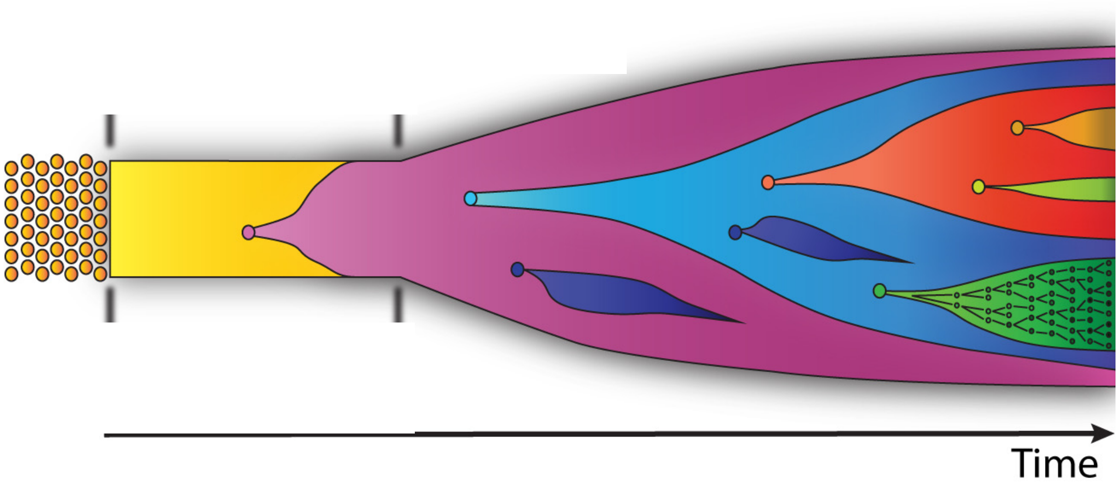
★ Lecture 1-Intro

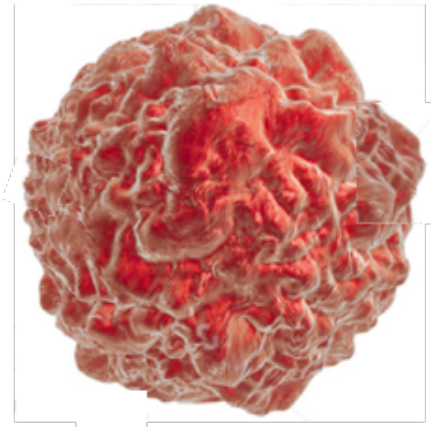
"Chapter 1"

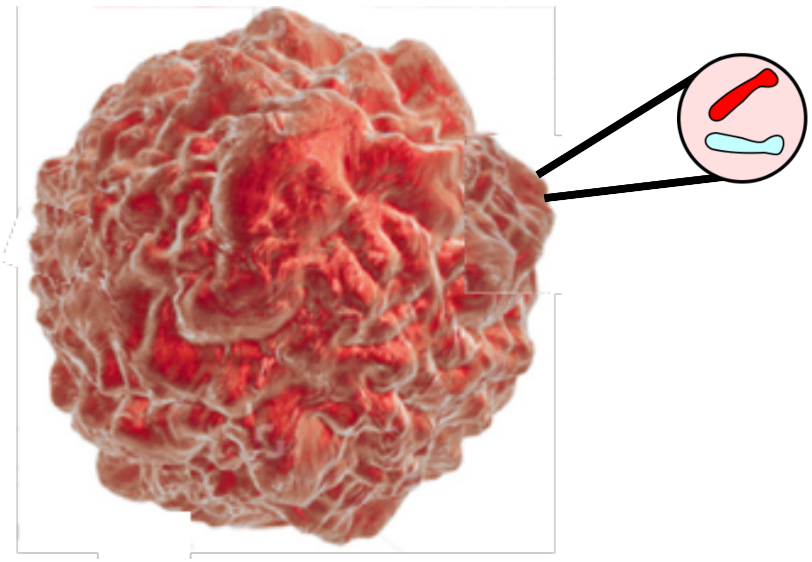

Karolinska Institutet Science Park
Science for Life Laboratory
SWEDISH ORPHAN BIOVITRUM

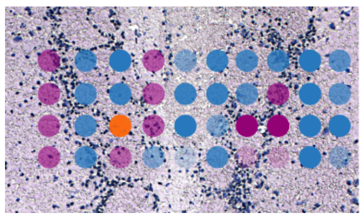

Time

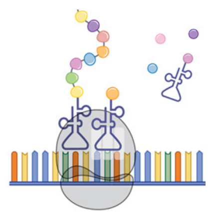


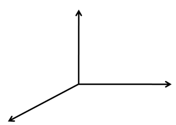

Time

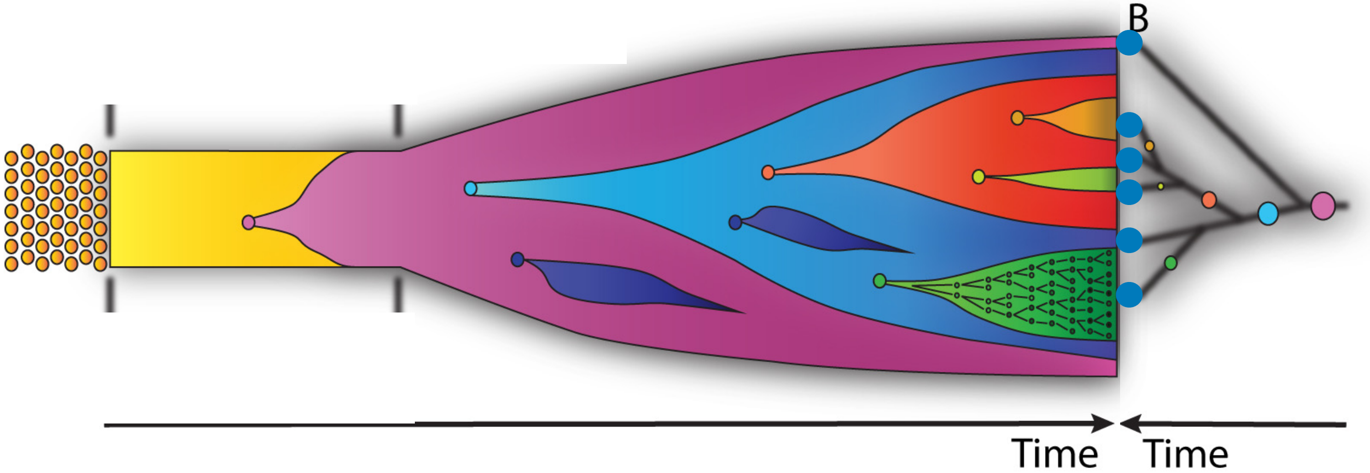


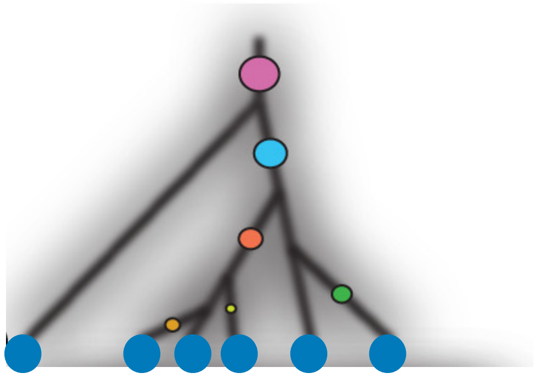


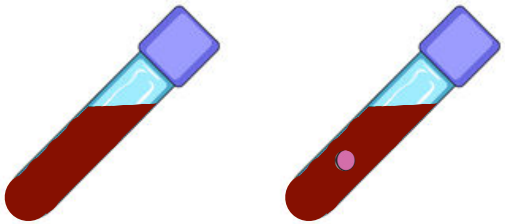


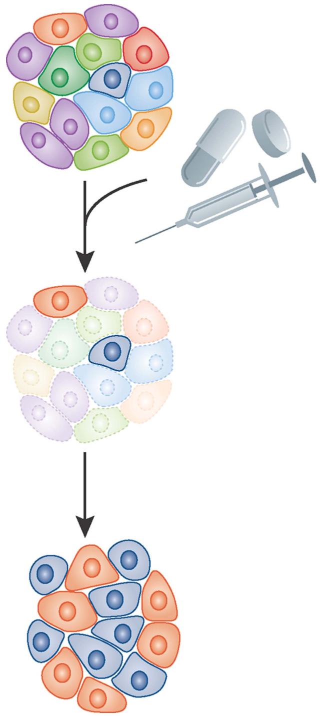


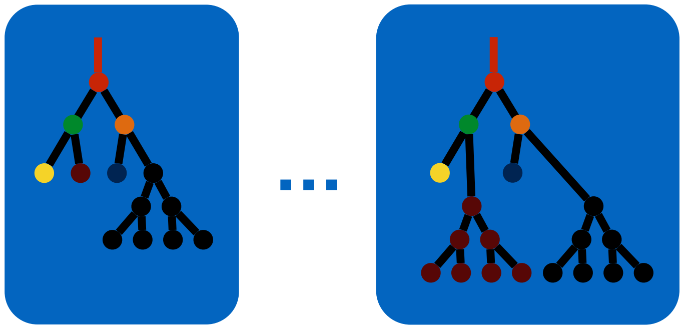



B
Time
Time

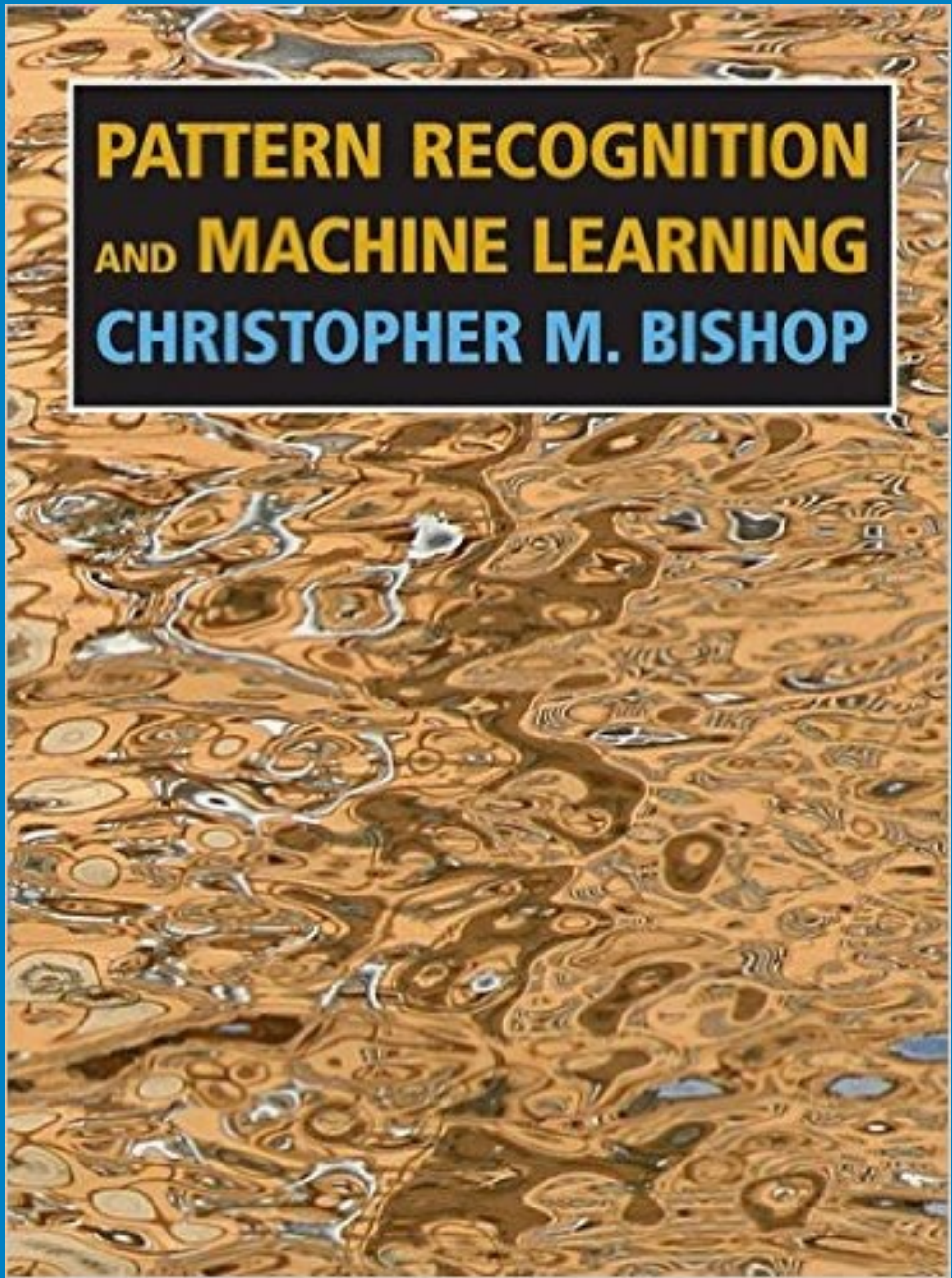








Oslo
Stockholm
Baltic Sea
Riga
North Sea
Denmark
Edinburgh
United Kingdom
Isle of Man
Dublin
Manchester
Liverpool
Hamburg
Poland
Berlin
Amsterdam
Netherlands
London
Brussels
Cologne
Germany
Belgium
Frankfurt
Prague
Luxembourg
Czechia
Paris
Slovakia
Munich
Vienna
Budapest
Austria
Switzerland
France
Slovenia
Zagreb
Milan
Croatia
Belgrade
Београд
Bosnia and Herzegovina
Serbia
Sarajevo
Bay of Biscay
Monaco
Italy
Andorra
Podgorica
Подгорица
Kosovo
Rome
Macedonia (FYROM)
Barcelona
Porto
Tirana
Albania
Madrid
Portugal
Valencia
Spain
Tyrrhenian Sea
Greece
Lisbon
Athens
Αθήνα
Seville
Granada
Algiers
Tunis
Málaga
Gibraltar



THE TEXT BOOK

- Key machine learning researcher
- Reasonably mathematical
- Well-written

CHANGES

* E and D level without problem solving

ADVICE FROM TWO YEARS AGO

- ★ Read Bishop's book
- ★ Go to all lectures. Read up. Start with the project on time.
- ★ Avoid the lectures, study Bishop from the beginning because in the other hand you will find yourself in a position that you will be running to catch up with the deadlines.

ADVICE & COMMENTS

- The assignments were thorough, so you actually had to understand the material to pass. The course contents were very interesting.
- The assignments takes at least three times longer than you think, start on time with everything.
- Don't underestimate the time it takes to finish the assignments Do the project and assignment as early as possible
- The problems are intentionally made so you cannot Google them…
- In addition, the lectures did not reflect at all on what we should go through in the assignments, so it was basically a matter of own work…
- Solving the assignments made me understand and made everything clear…

or

- ★ Homework (2) – Individual, handed in and graded
  - ★ First out now, or next week, in December 1
  - ★ Second out November 25, in Dec 23
- ★ Project (1) – Collaborative, handed in and graded
  - ★ Out November 20, in Jan 15
  - ★ You define the groups of four (based on merit)

EXAMINATION - APPROXIMATE DATES

1. Use latex, you will have to use it later anyway (this is a recommendation)

2. Always include your name in the file with the solutions

3. Make each step in a derivation explicit

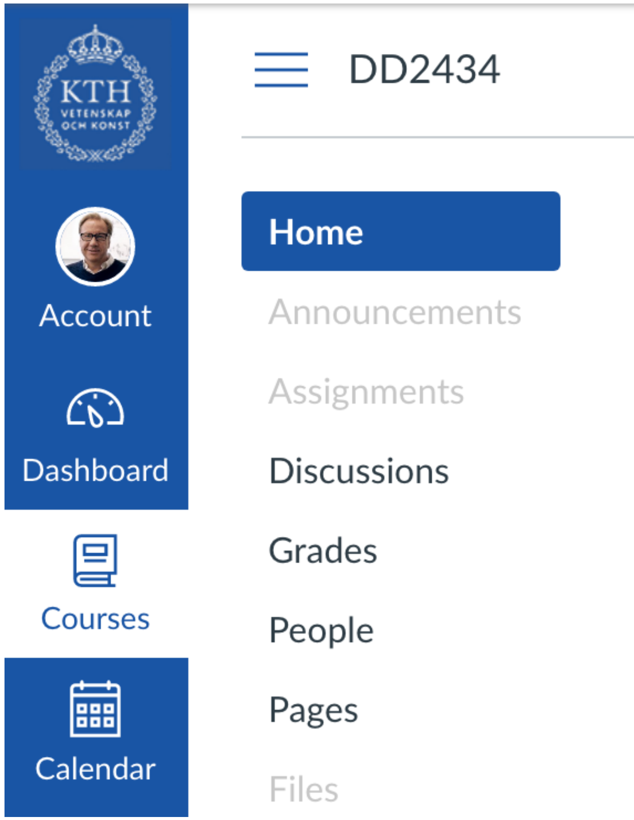
CONCERNING SOLUTIONS

HEDVIGIAN GRADING WITH POSSIBLE CHANGES

Grading

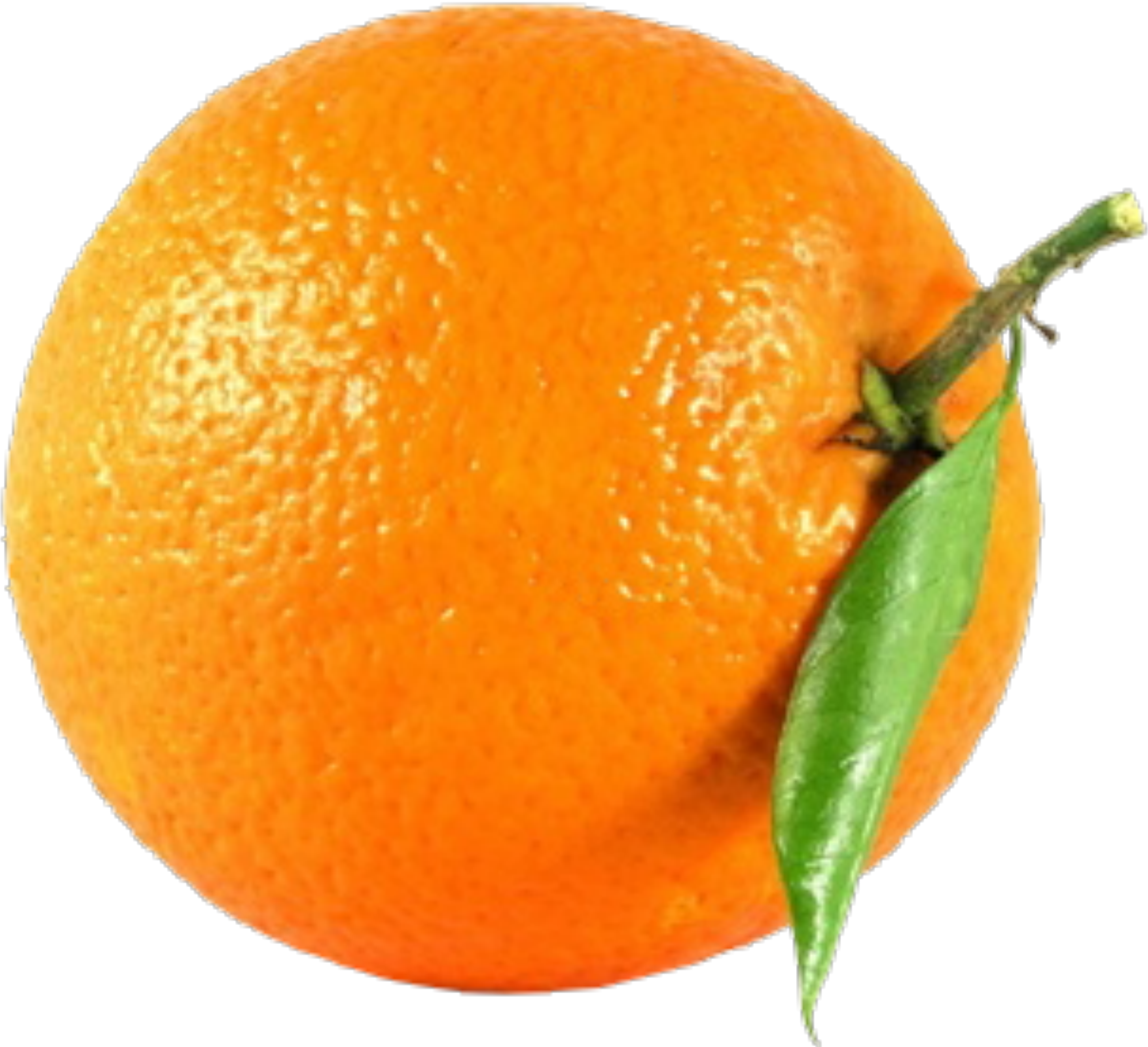
The course grade is the weighted average of the assignment grade and the project grade, according to the following:

Assignment \ Project	A	B	C	D	E
A	A	A	B	B	C
B	B	B	B	C	C
C	B	C	C	C	D
D	C	C	D	D	D
E	C	D	D	E	E

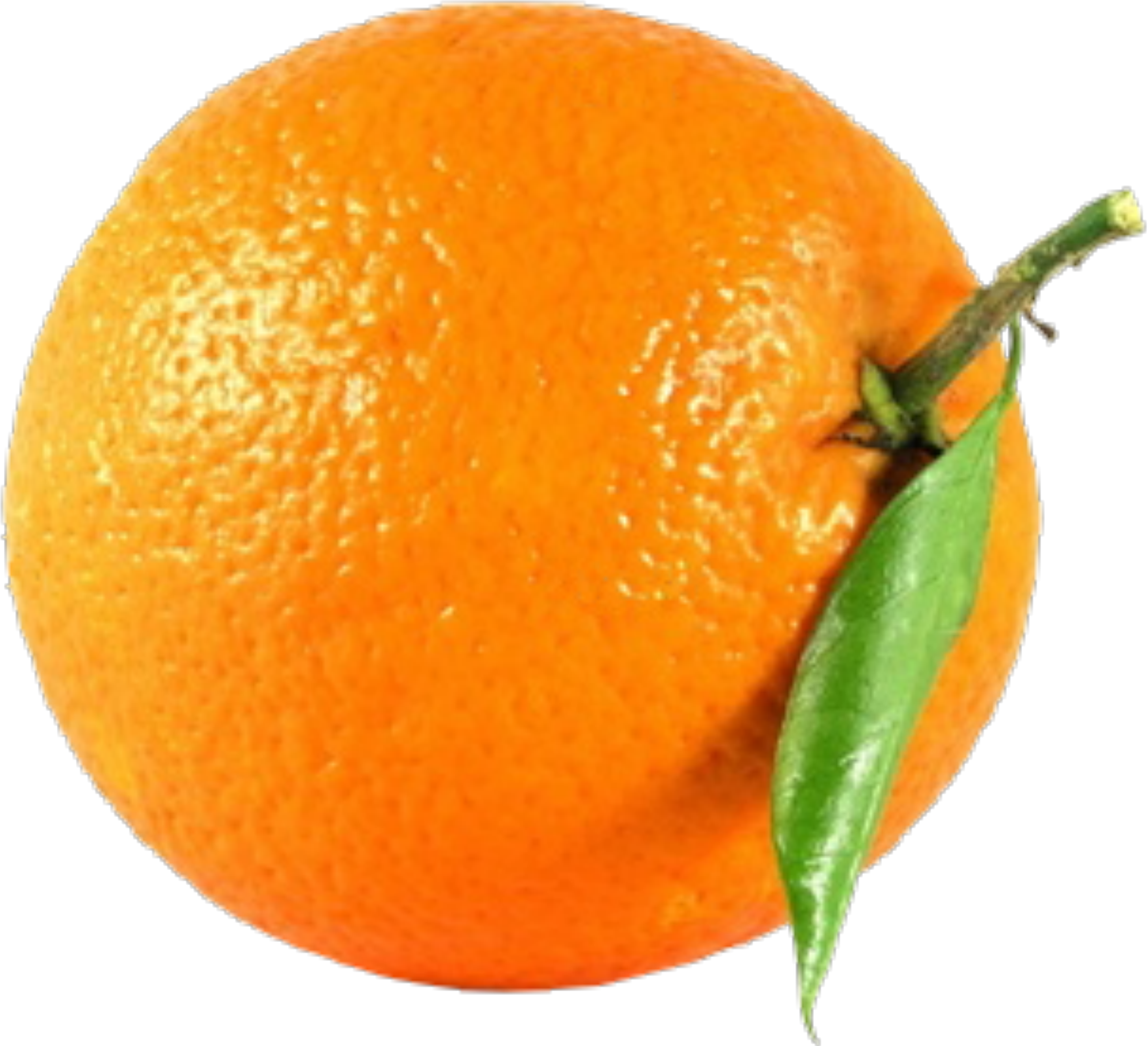


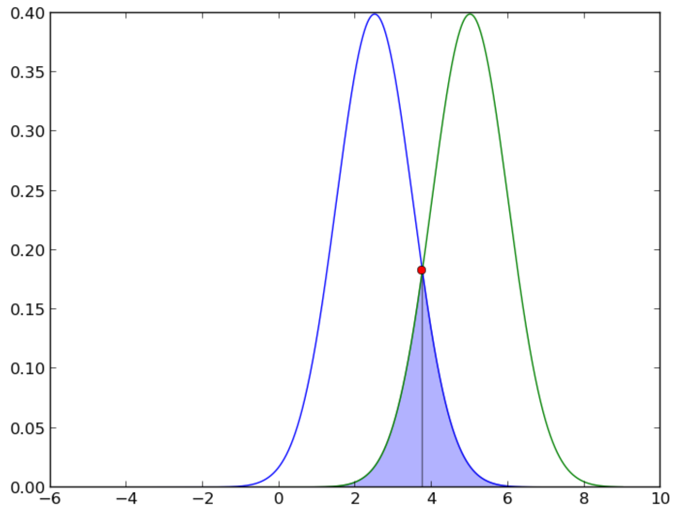
INTERACTION

- Lectures
- Solutions: CANVAS.
- My slides & notes: Slack
- Assignment 2: Slack

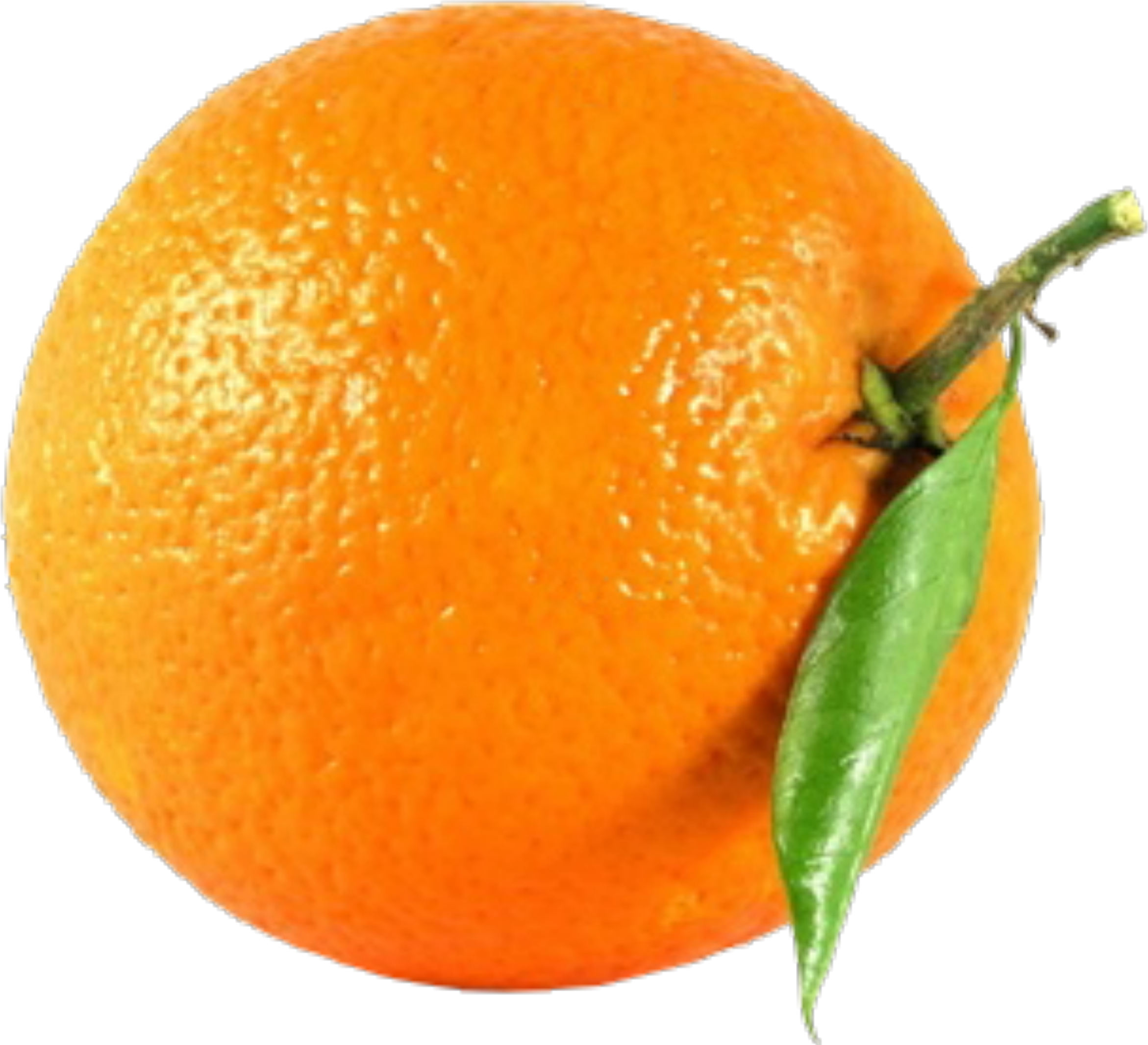


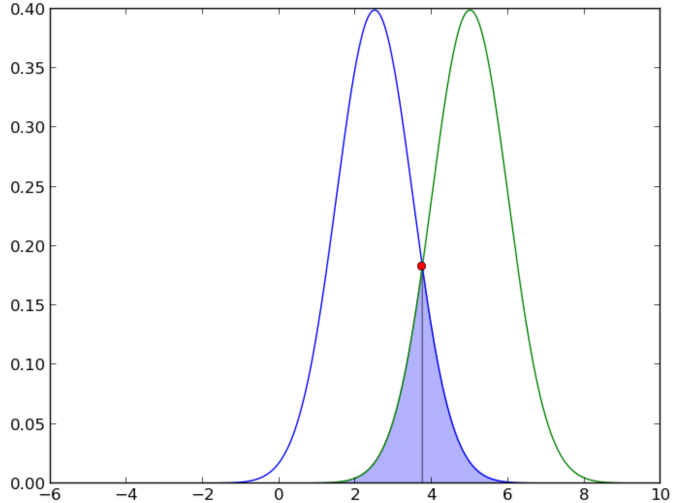
SMALL OR LARGE

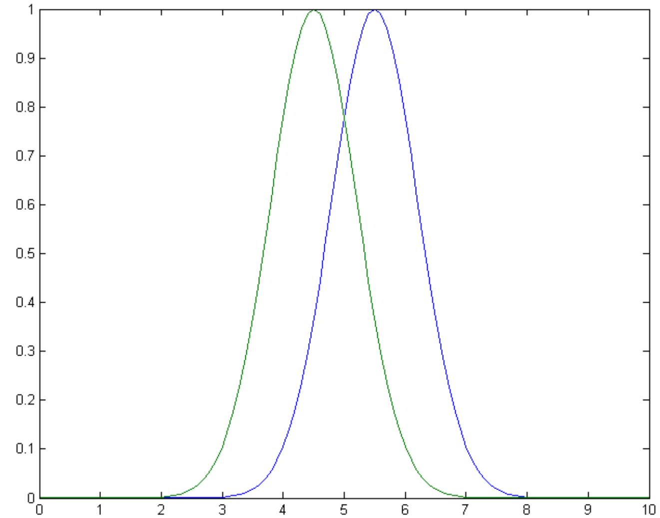




SMALL OR LARGE

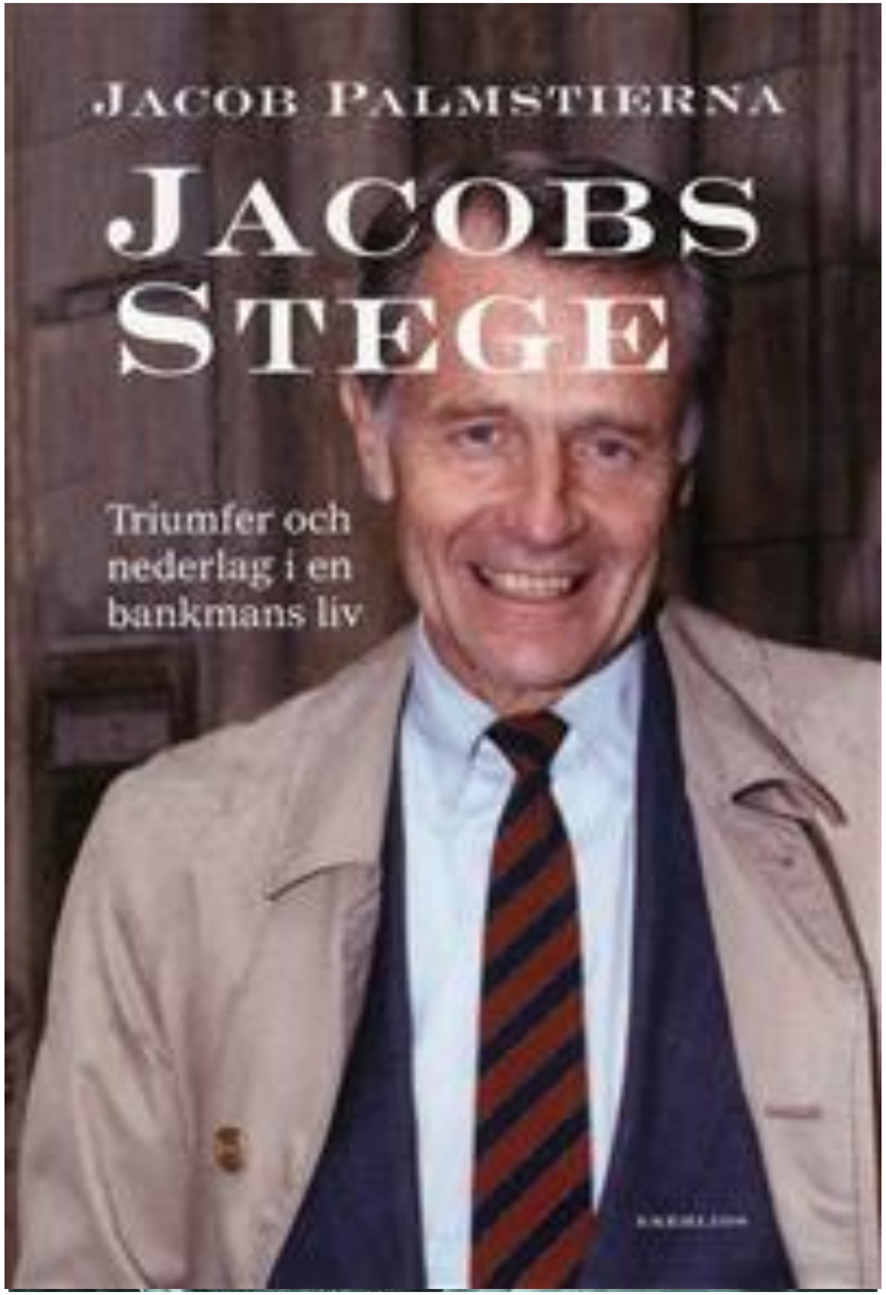


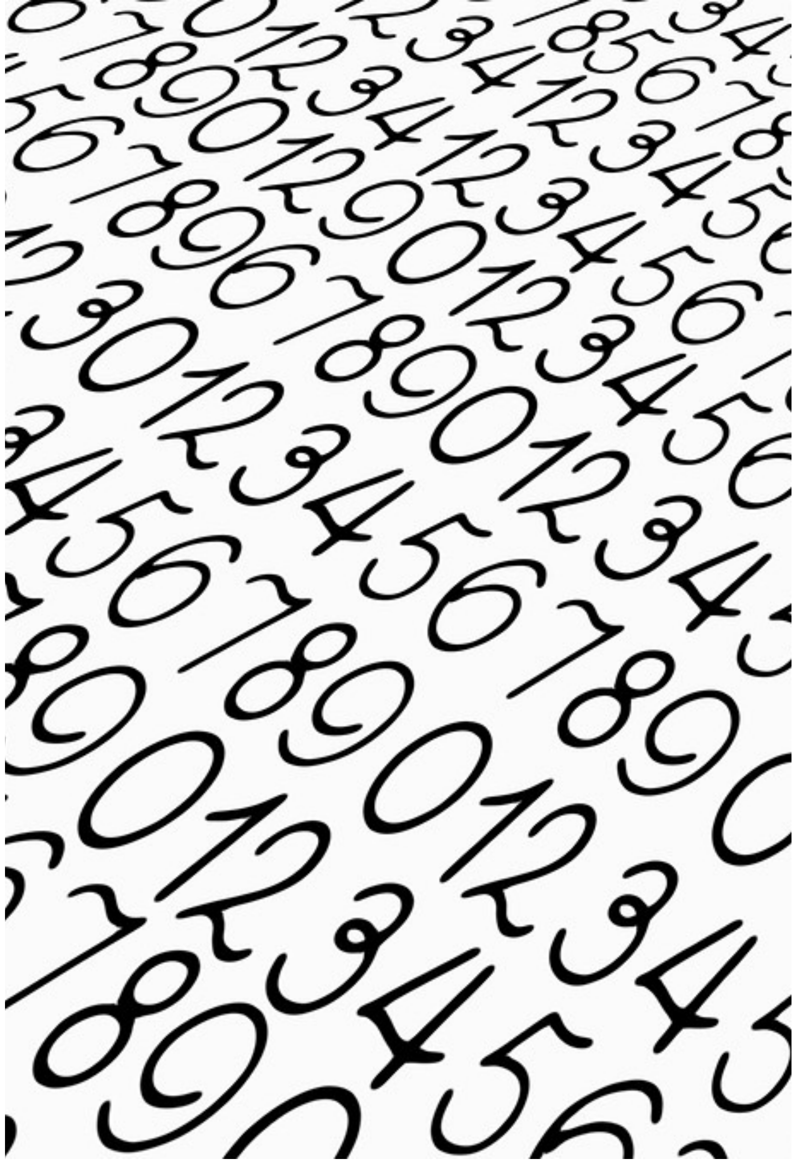




SMALL OR LARGE

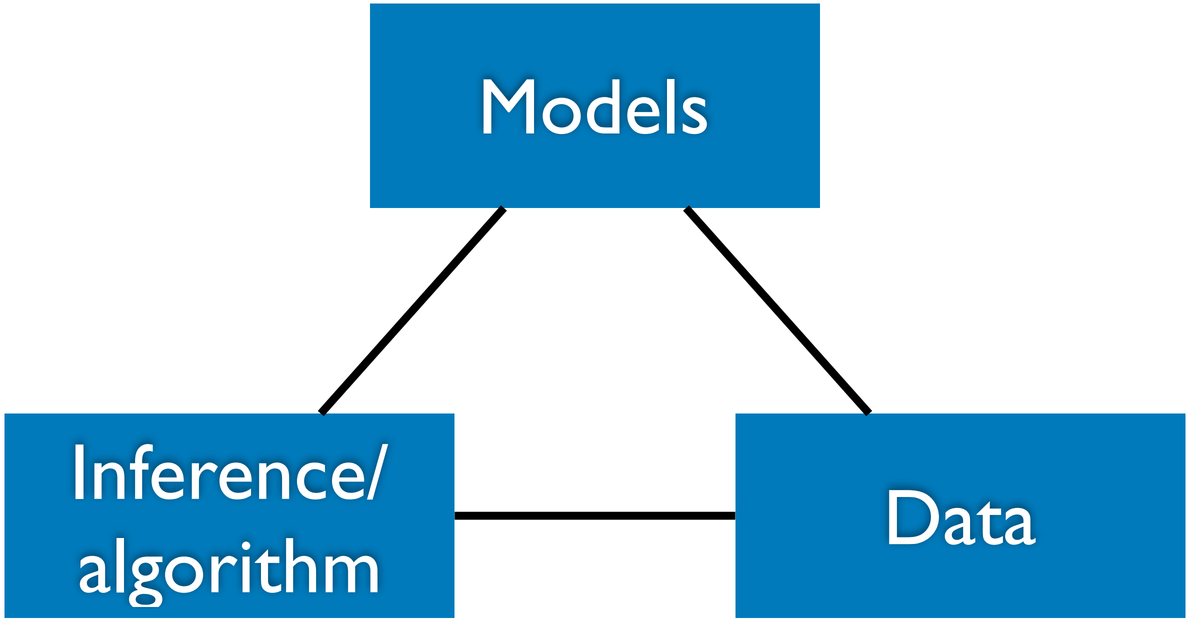
TAKING DECISIONS



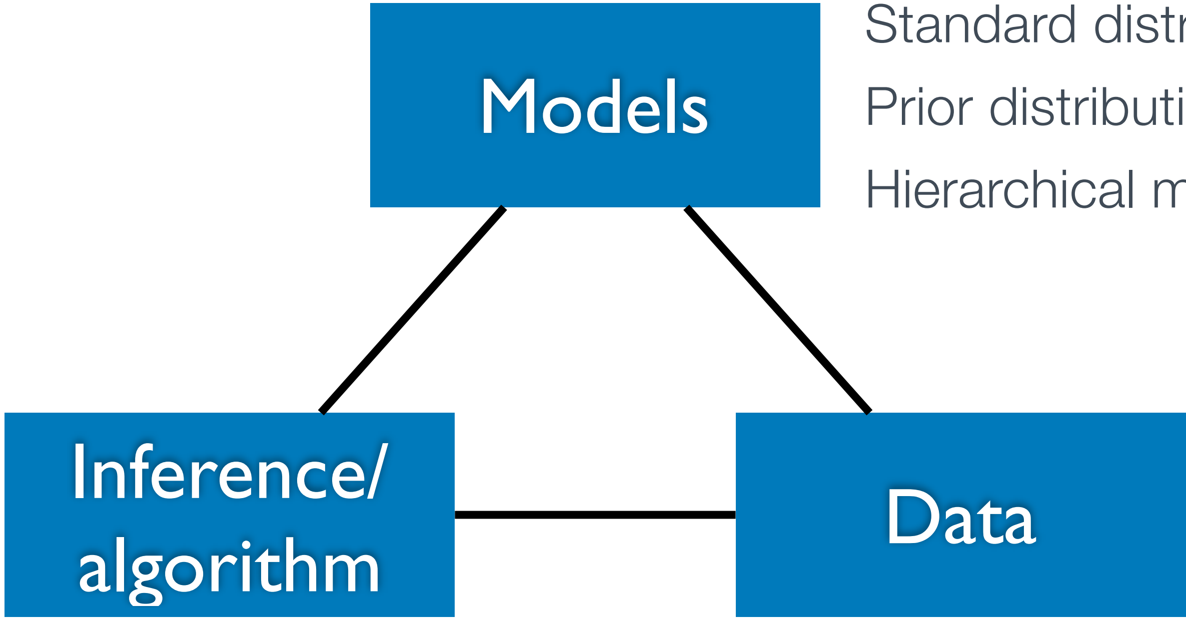


TAKING DECISIONS IN THE PRESENCE OF UNCERTAINTY

MACHINE LEARNING



MACHINE LEARNING



Standard distributions (MVN)
Prior distributions
Hierarchical models

Maximum likelihood
Bayesian

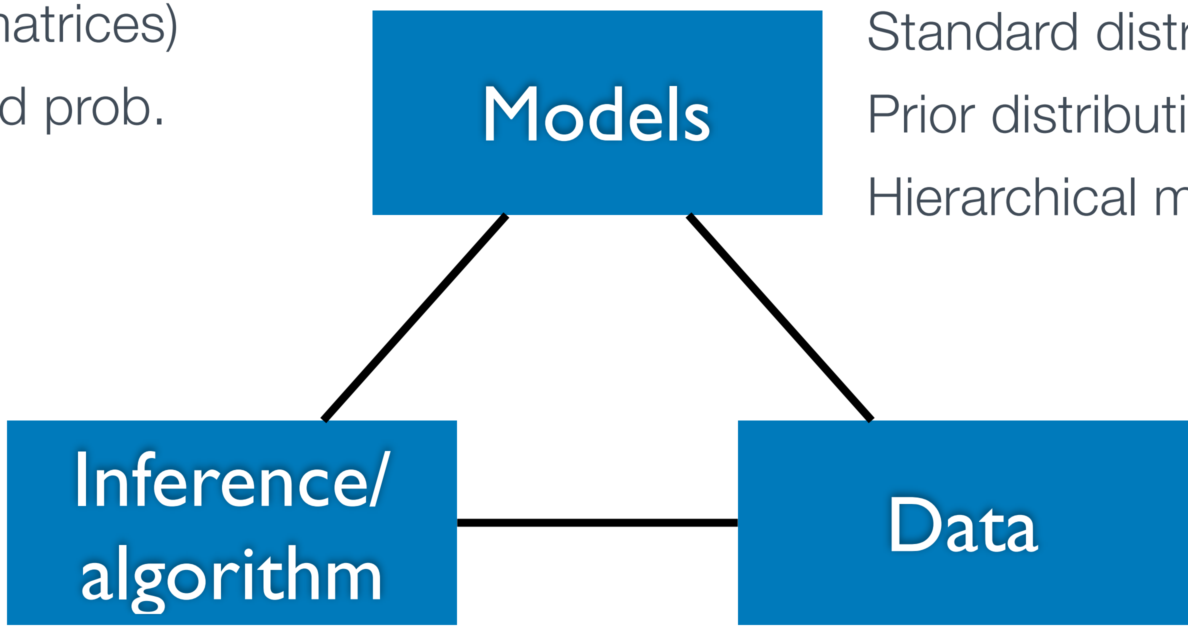
Likelihood for HMMs
Expectation Maximization
Variational Bayes

Size
Errors

MACHINE LEARNING

Have a look at

- Linear algebra (matrices)
- Previous stat. and prob. courses
- Chapter 2.



Standard distributions (MVN)

Prior distributions

Hierarchical models

Maximum likelihood

Bayesian

Likelihood for HMMs

Expectation Maximization

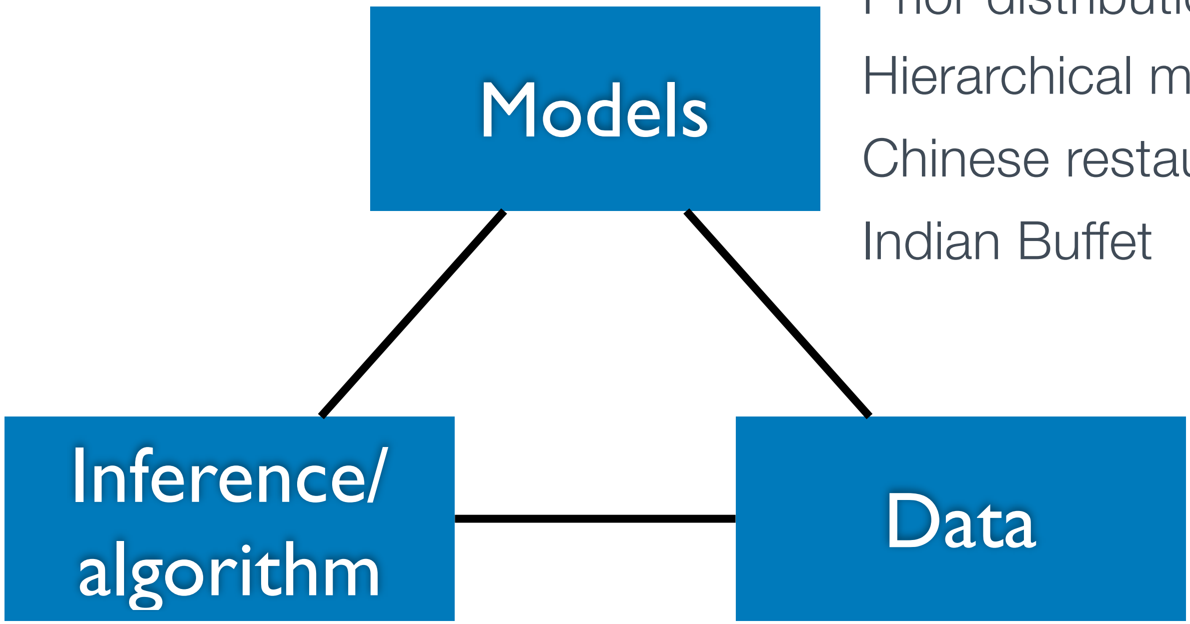
Variational Bayes

Size

Errors

MACHINE LEARNING DD2447

Standard distributions
(Categorical, Dirichlet)
Prior distributions
Hierarchical models
Chinese restaurant
Indian Buffet



Maximum likelihood
Bayesian

Simulation
Sampling

Size
Errors

SOME THOUGHTS ON MODELING

- All models are wrong, but some are useful.
  - Models are what we call the lies we are used to
- There are no model free approaches!
  - use the term assumption instead
- Using models is a way to make assumptions explicit.
- Bayesian is a non-deterministic logic.

SOME STUFF I EXPECT YOU TO KNOW

- Supervised learning
- Unsupervised learning
- Training & testing
- Probability

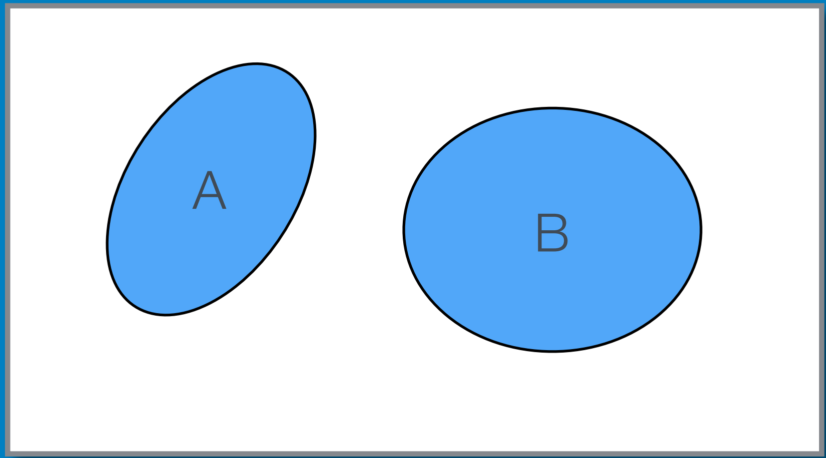
PRODUCT RULE: CONDITIONING

$$p(x, y) = p(y)p(x|y)$$

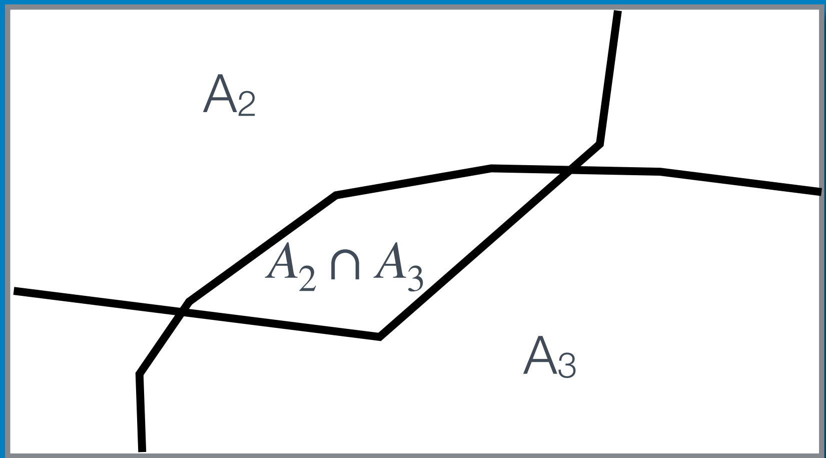
or

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

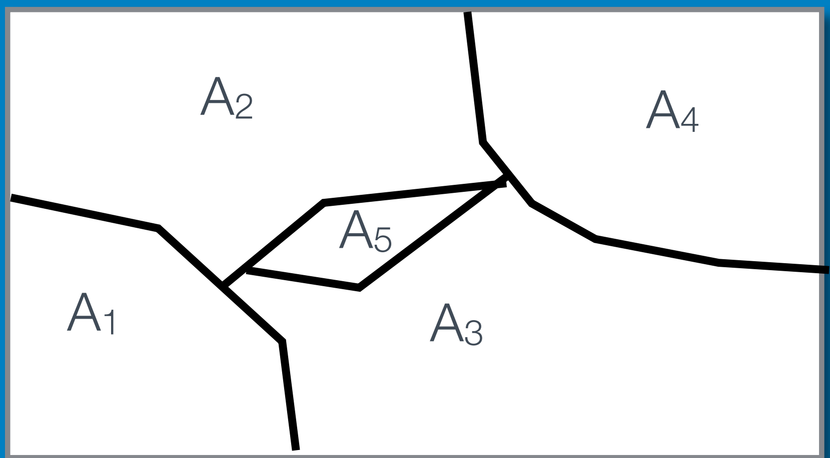
Exclusive



Exhaustive



Exclusive & exhaustive



EXCLUSIVE & EXHAUSTIVE

- Exclusive

$$p(A \text{ or } B) = p(A) + P(B)$$

- Exclusive & exhaustive

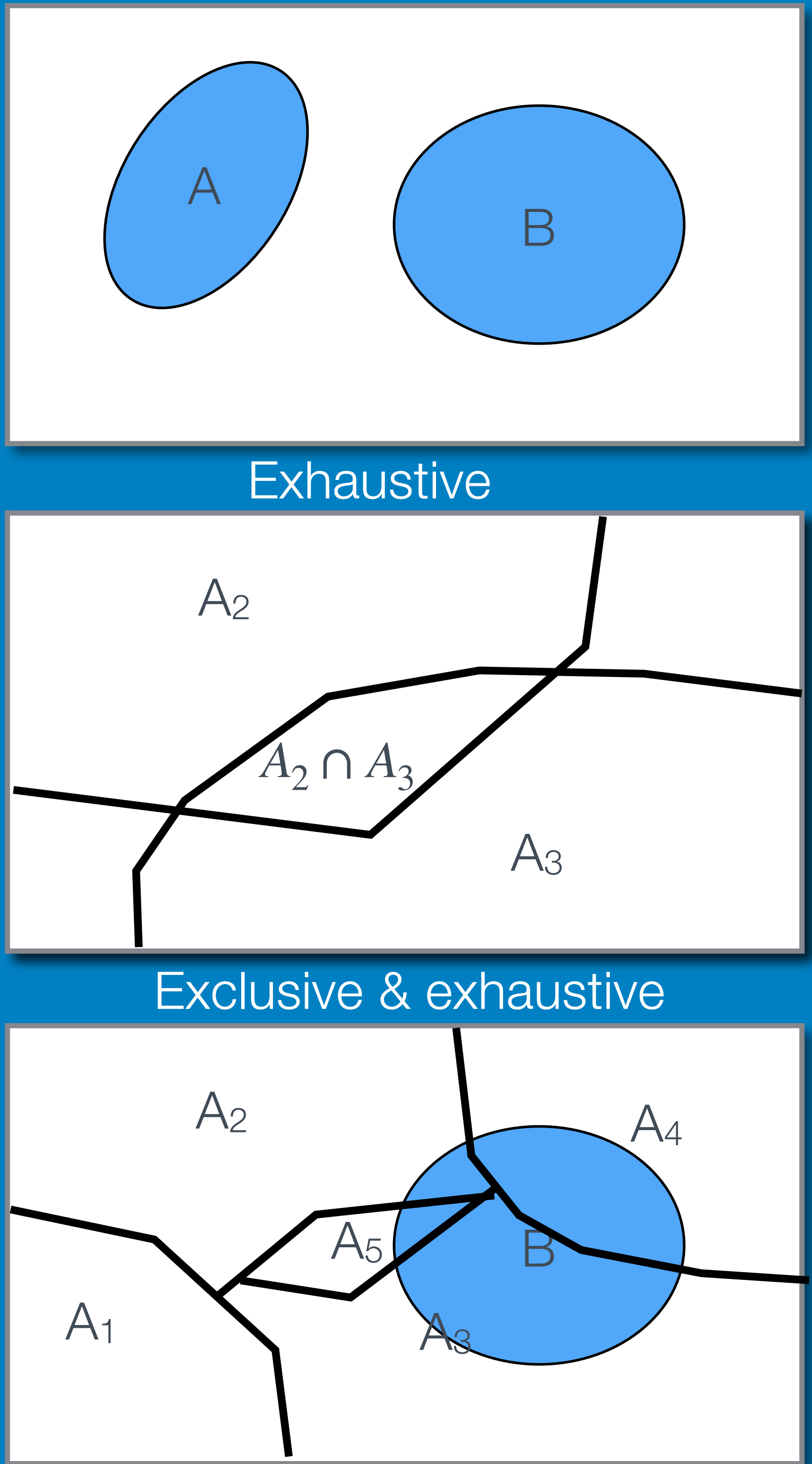
$$\sum_i p(A_i) = 1$$

SUM RULE: EXCLUSIVE & EXHAUSTIVE

Exclusive

Exhaustive

Exclusive & exhaustive



- Exclusive

$$p(A \text{ or } B) = p(A) + P(B)$$

- Exclusive & exhaustive

$$p(B) = \sum_i p(B, A_i) = \sum_i p(A_i) p(B|A_i)$$

BAYES

$$p(M \mid X) = \frac{p(X, M)}{p(X)} = \frac{p(X \mid M)p(M)}{\sum_M p(X \mid M)p(M)}$$

BAYESIAN

Fair (F)

$$P(i|F) = \frac{1}{6}$$

Used 99%

Biased/loaded (B)

$$P(6|B) = \frac{1}{2}$$

$$P(6|B) = \frac{1}{10}, \forall i \in [5]$$

Used 1%

- A is the event 6,6,6

BAYESIAN

Fair (F)

$$P(i|F) = \frac{1}{6}$$

Used 99%

Biased/loaded (B)

$$P(6|B) = \frac{1}{2}$$

$$P(6|B) = \frac{1}{10}, \forall i \in [5]$$

Used 1%

- A is the event 6,6,6

- We get $$P(M|A) = \frac{P(A|M)P(M)}{P(A)}$$

The same for F & B

$$P(A|F)P(F) = \frac{1}{6}^{3} * 0.99$$

$$P(A|B)P(B) = \frac{1}{2}^{3} * 0.01$$

BAYESIAN

Fair (F)

$$P(i|F) = \frac{1}{6}$$

Used 99%

Biased/loaded (B)

$$P(6|B) = \frac{1}{2}$$

$$P(6|B) = \frac{1}{10}, \forall i \in [5]$$

Used 1%

- A is the event 6,6,6

- We get $$P(M|A) = \frac{P(A|M)P(M)}{P(A)}$$

The same for F & B

$$P(A|F)P(F) = \frac{1}{6}^3 * 0.99 \quad > \quad P(A|B)P(B) = \frac{1}{2}^3 * 0.01$$

MAXIMUM LIKELIHOOD (ML) AND POSTERIOR PREDICTIVE

- ML
  - Estimate θ_{ML} from training data D and then use

$$y(x, \theta_{ML})$$

- Bayesian
  - Estimate a posterior distribution over θ based on D and then use

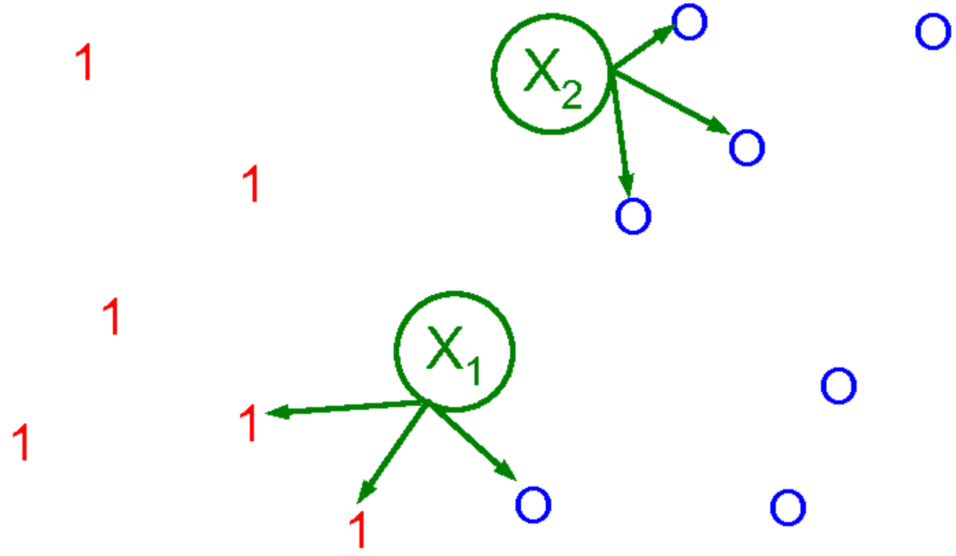
$$\int y(x, \theta) p(\theta \mid \mathscr{D}) d\theta$$

PARAMETRIC VS NON-PARAMETRIC

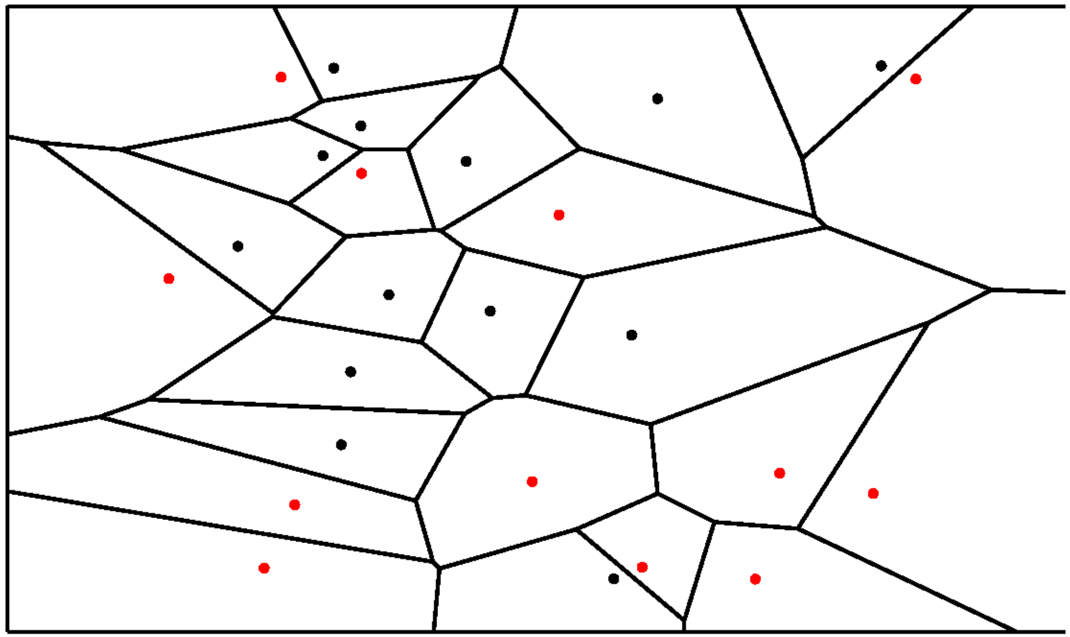
- Constant # parameters – parametric model (any distribution)
- Representation grows with data – non-parametric model

K-NN $$p(y=c|\boldsymbol{x},\mathcal{D},K)=\frac{1}{K}\sum_{n\in N_K(\boldsymbol{x},\mathcal{D})} I(y_n=c)$$

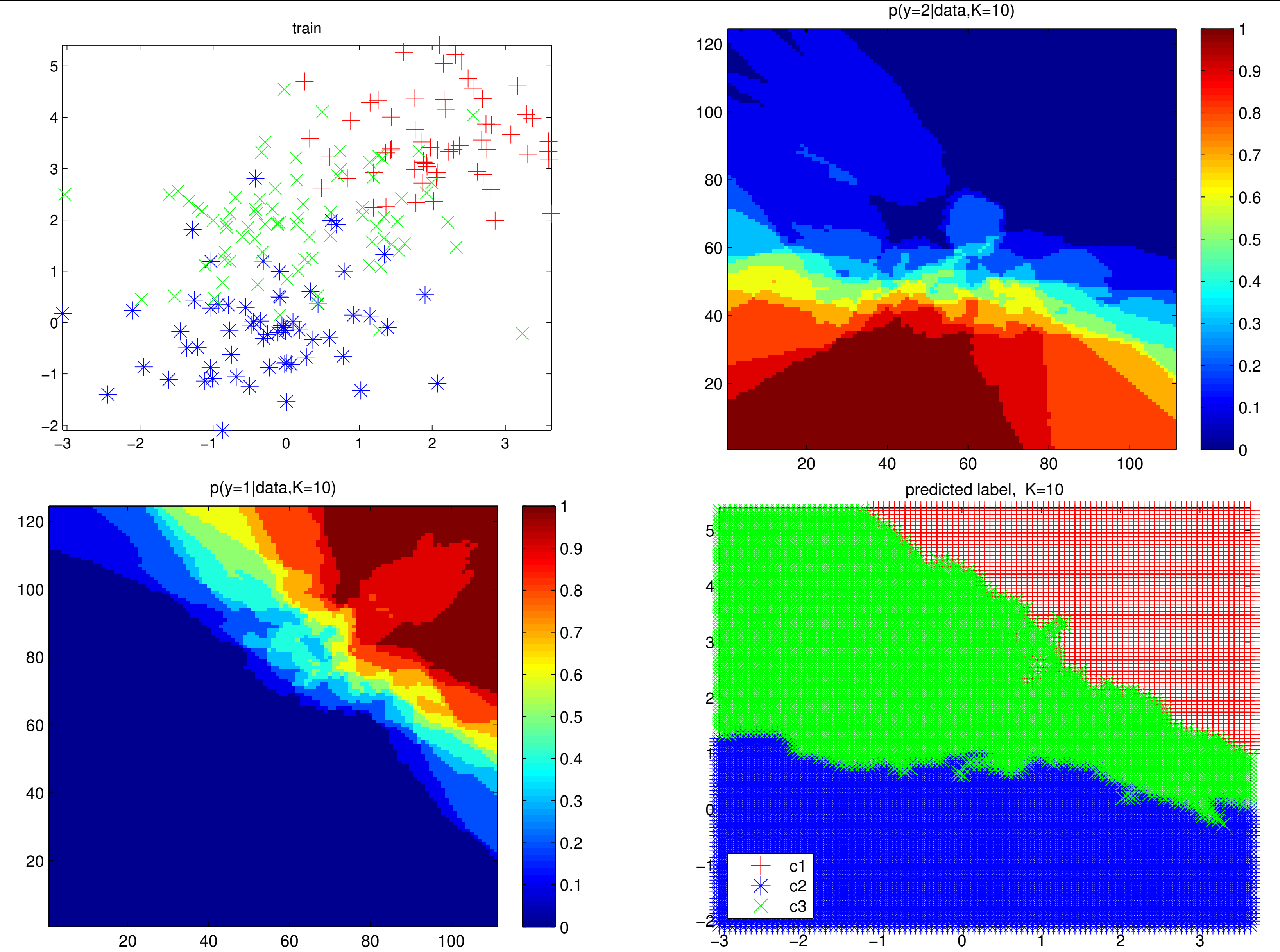
3-NN



1-NN



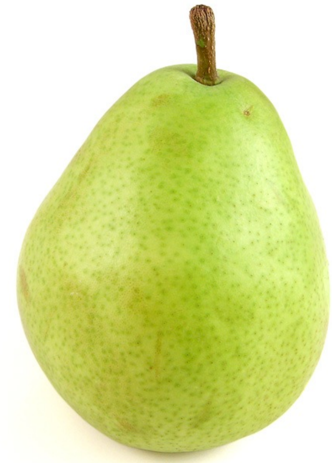
K-NEAREST NEIGHBOUR (K-NN)



K-NEAREST NEIGHBOUR

UNSUPERVISED LEARNING

- We do not have any correct answer
- Find classes or groups

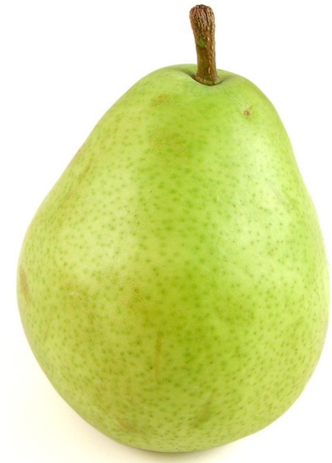


We have the answer means yes otherwise no

SUPERVISED LEARNING

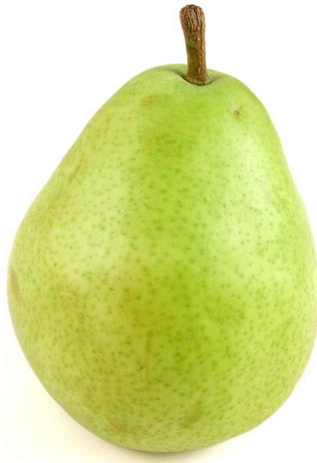
SOME STUFF I EXPECT YOU TO KNOW

- Supervised learning
  - $D=\{(\mathbf{x}_i, y_i)\}$
  - y_i response variable (output variable)
  - **x**_i features (input variables)
  - classification & regression
- Unsupervised learning
  - find the right y_i's, or
  - find the right dependencies between the variables of $\mathbf{x}_i$

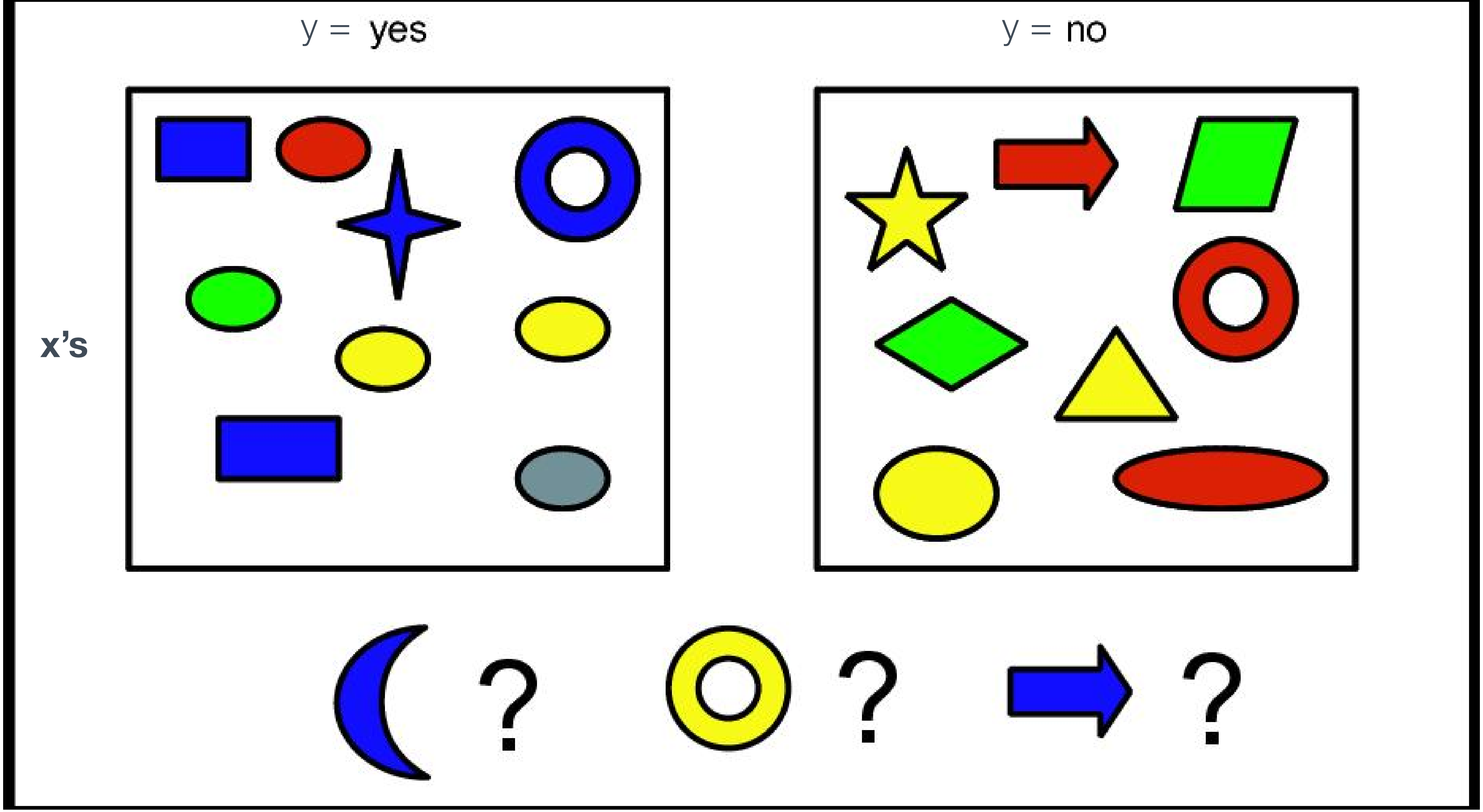

We have the answer means yes otherwise no

BINARY CLASSIFICATION

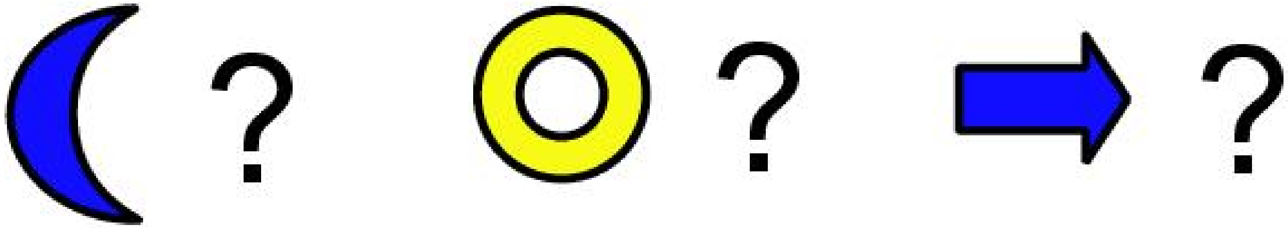


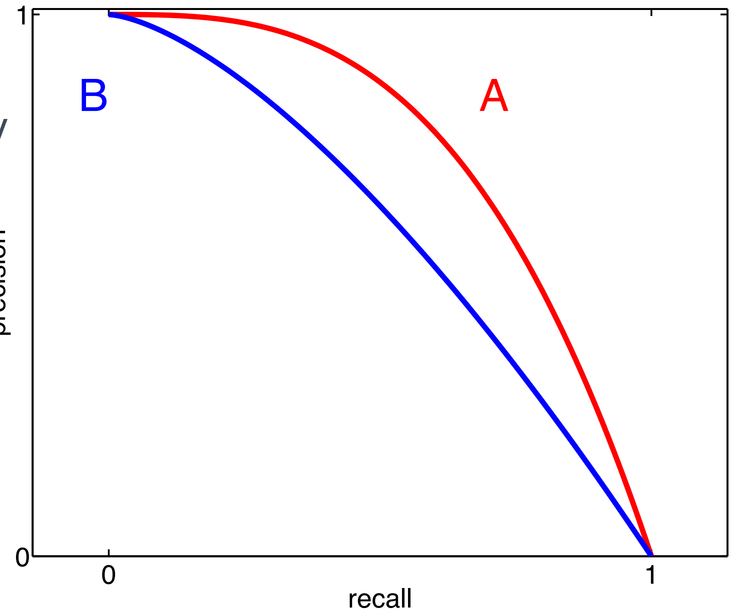
★ means apple, means pear, otherwise other

CATEGORICAL CLASSIFICATION



CLASSIFICATION





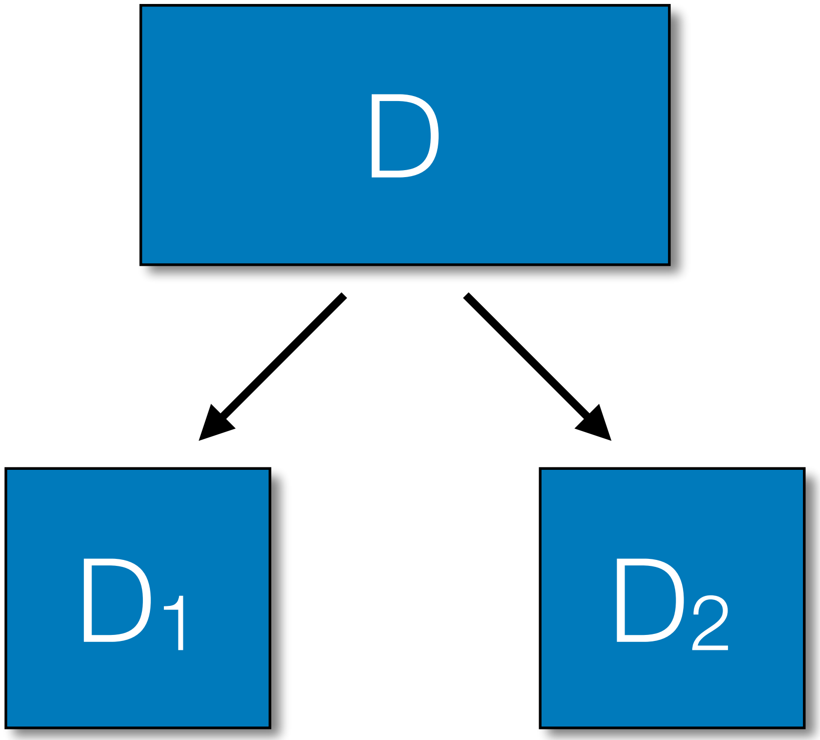
PPV

TPR, sensitivity

The probability of an answer

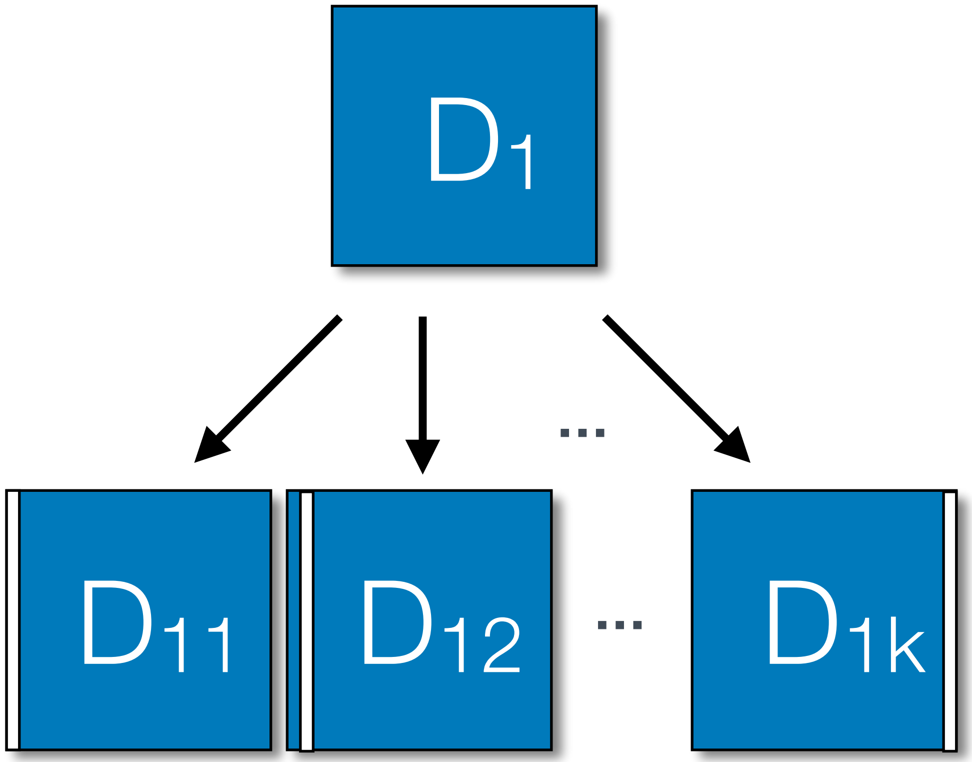
$$p(\text{ yes } \mid \text{blue moon}, \mathcal{D})$$

PROBABILISTIC PREDICTION



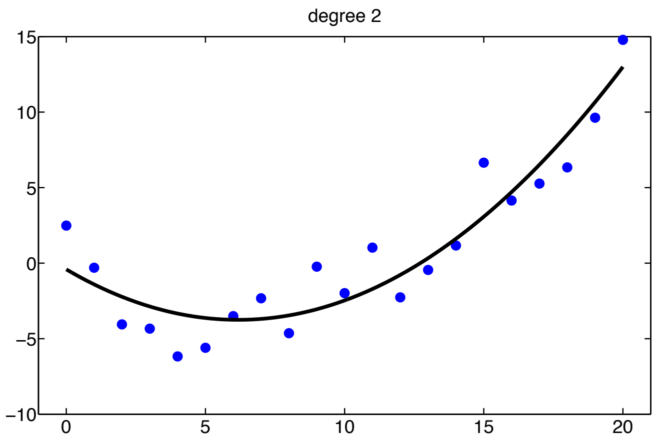
TRAINING – TEST

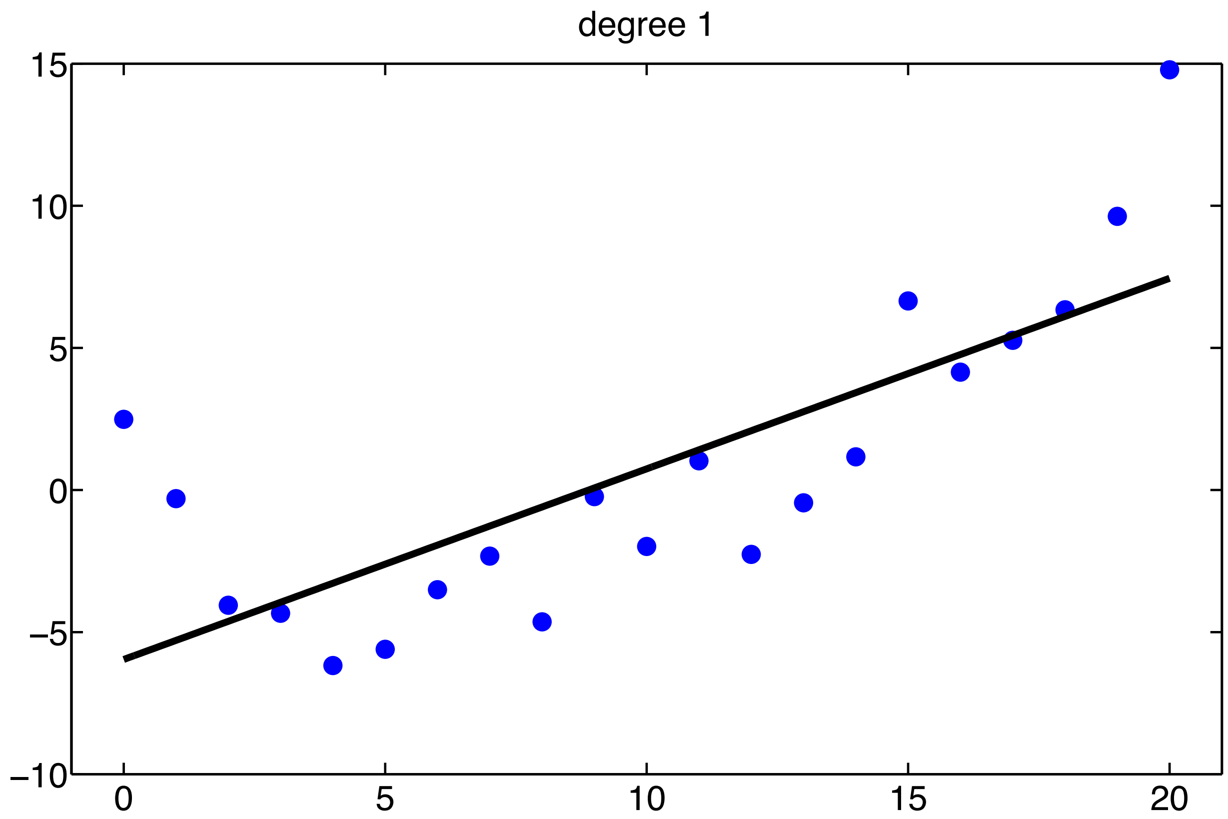
- Split data D into
  - training D_1
  - test D_2
- Use misclassification rate



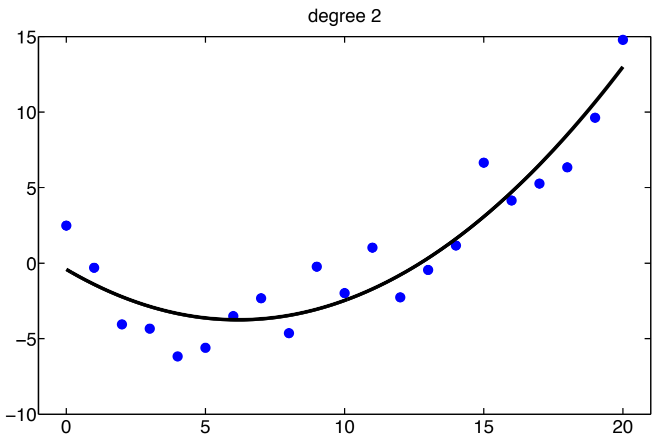
CROSS-VALIDATION

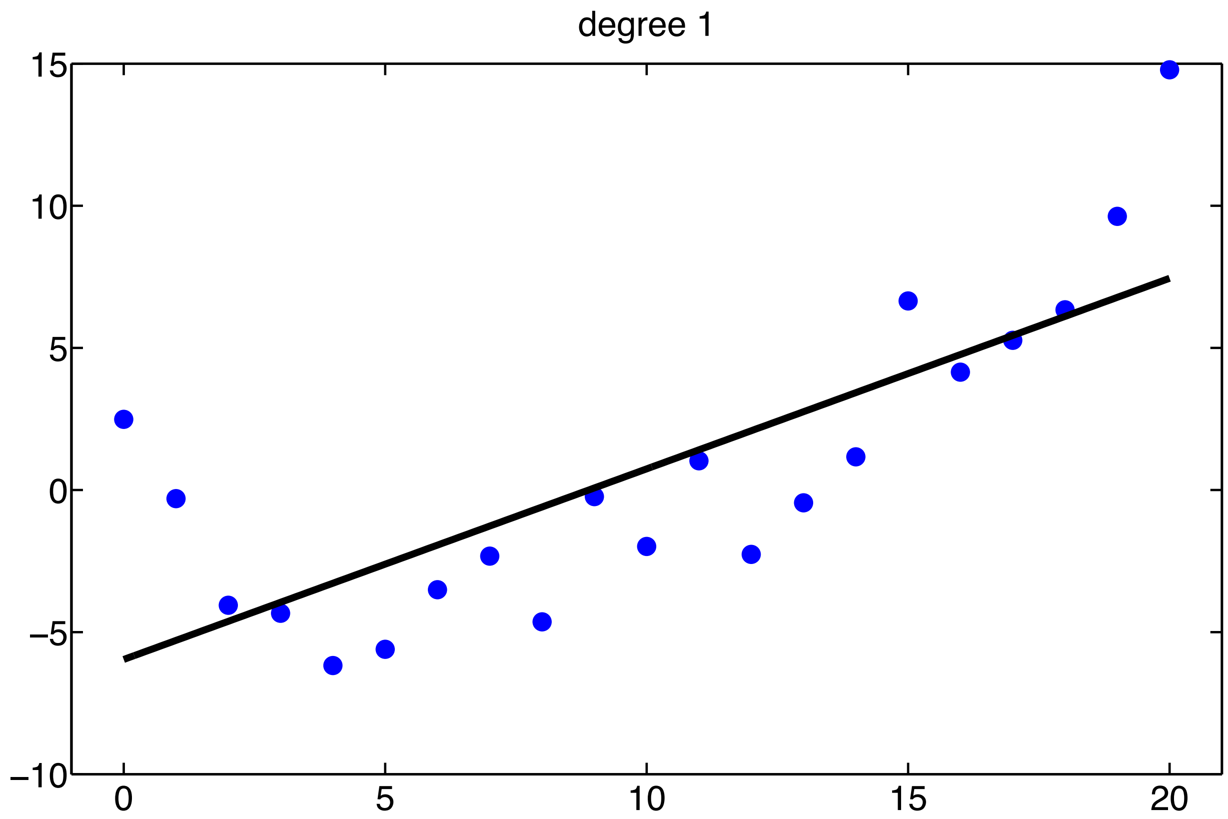
- Leave-one-out
  - let $D_{1i} = D_1 \setminus x_i$
  - test on x_i
- Use misclassification rate
- Redundancy, overlap





REGRESSION

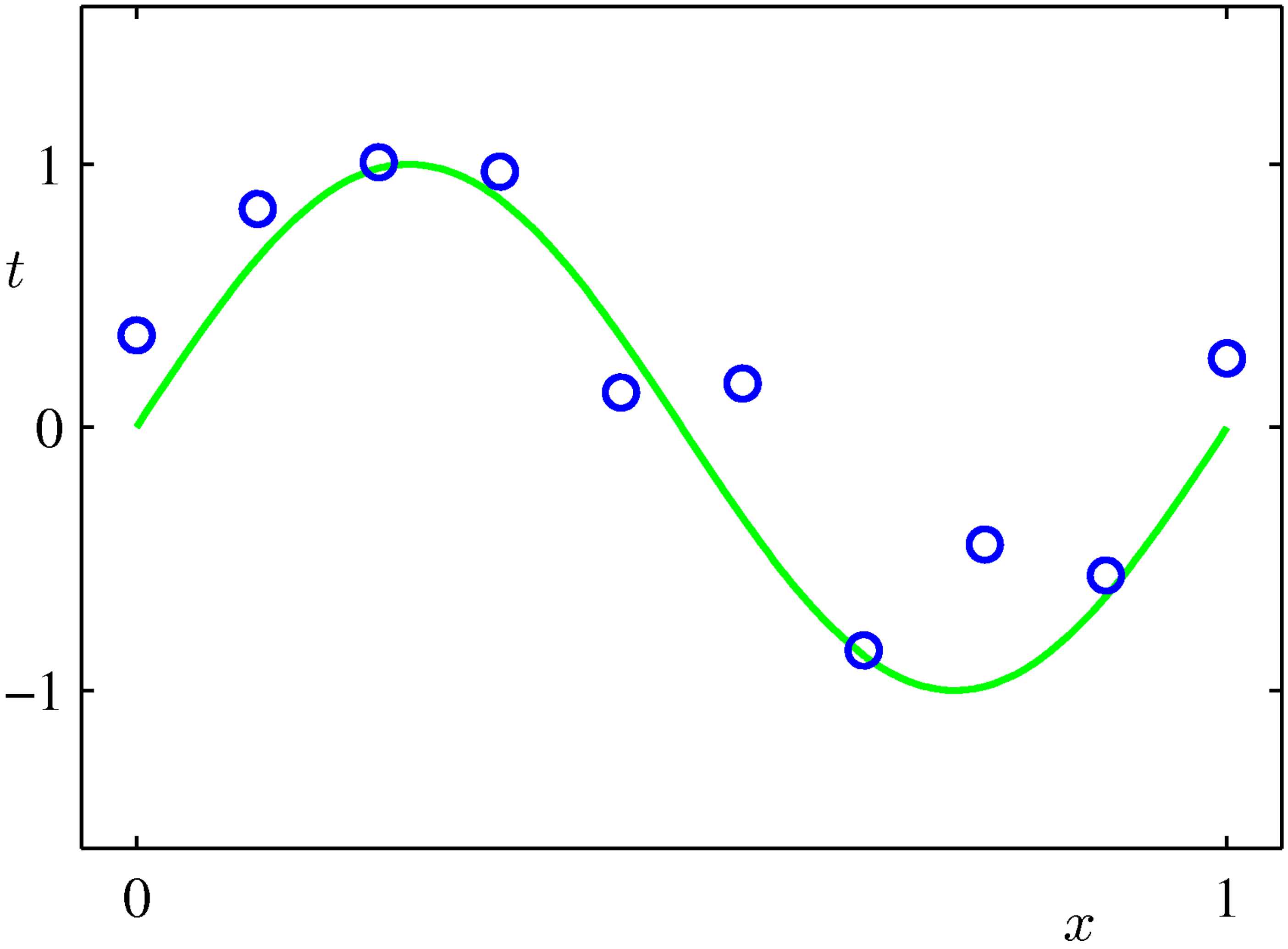




- Size
- Floor
- Location

REGRESSION

EX POLYNOMIAL FITTING - THE DATA

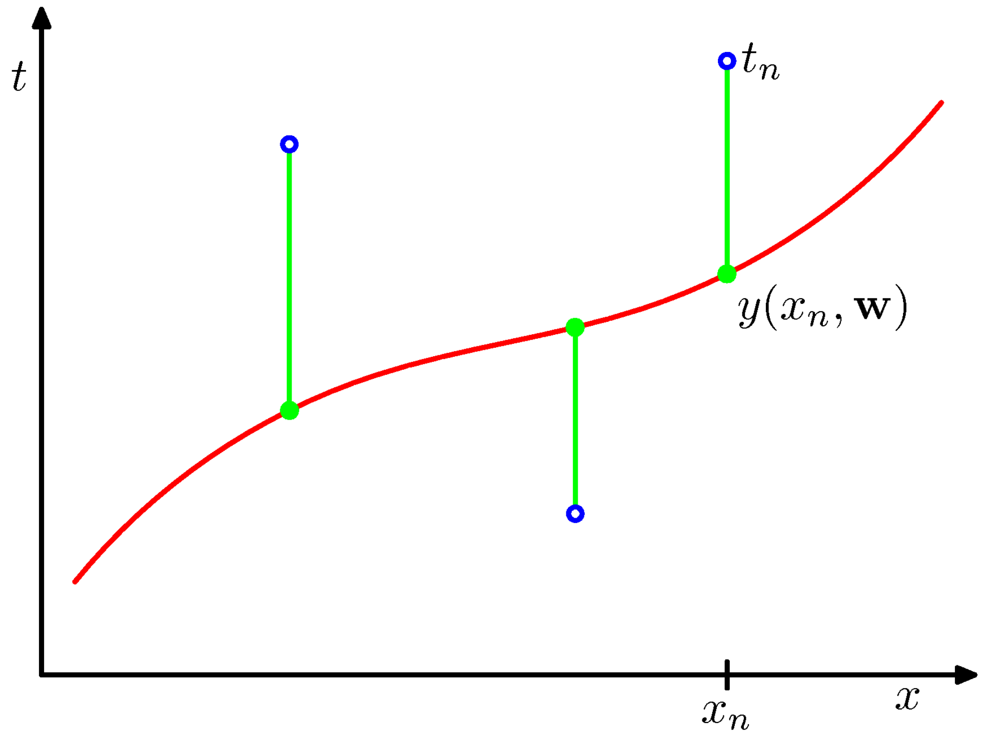


$\sin(2\pi x)$

with additive noise

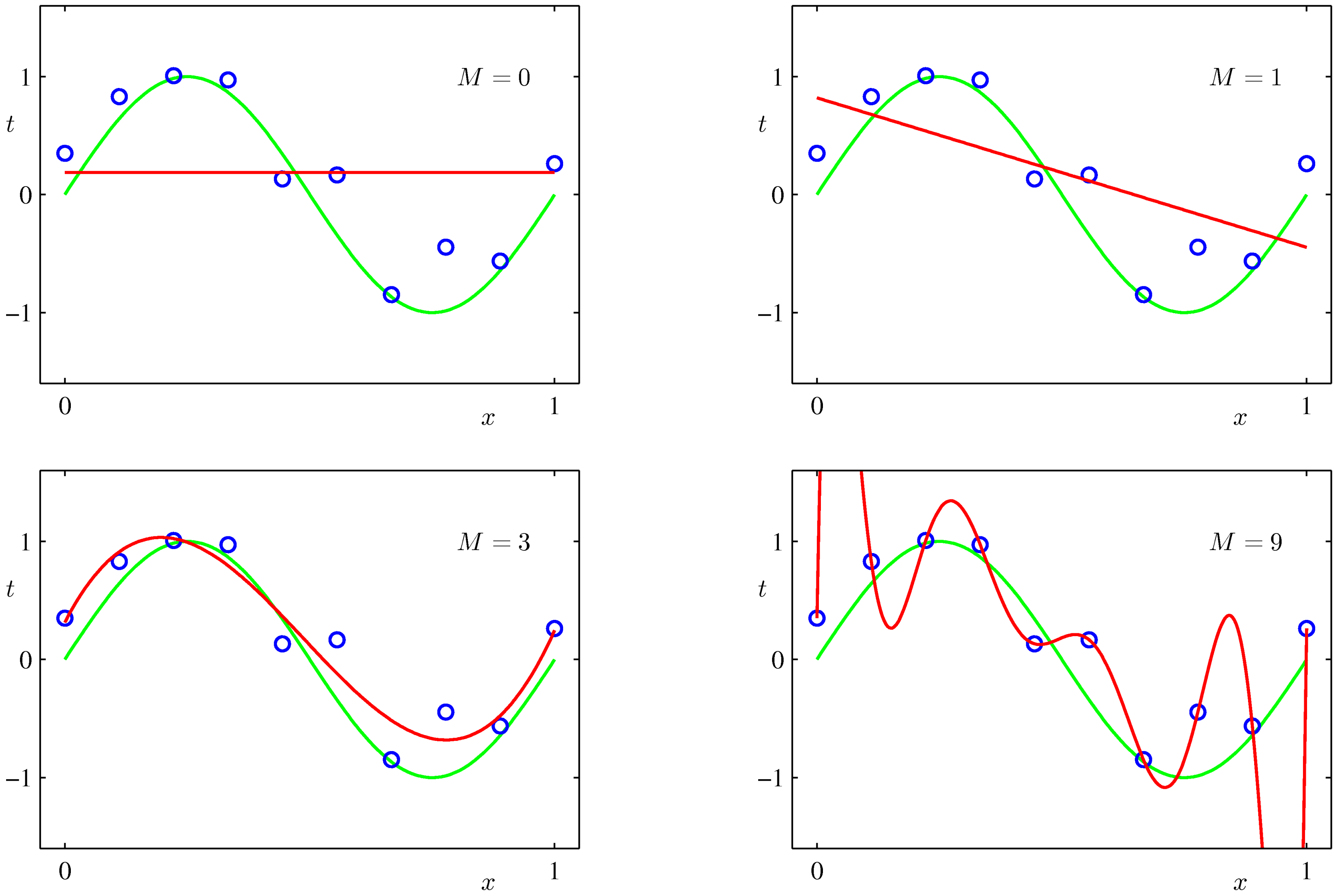
MEASURING ERROR

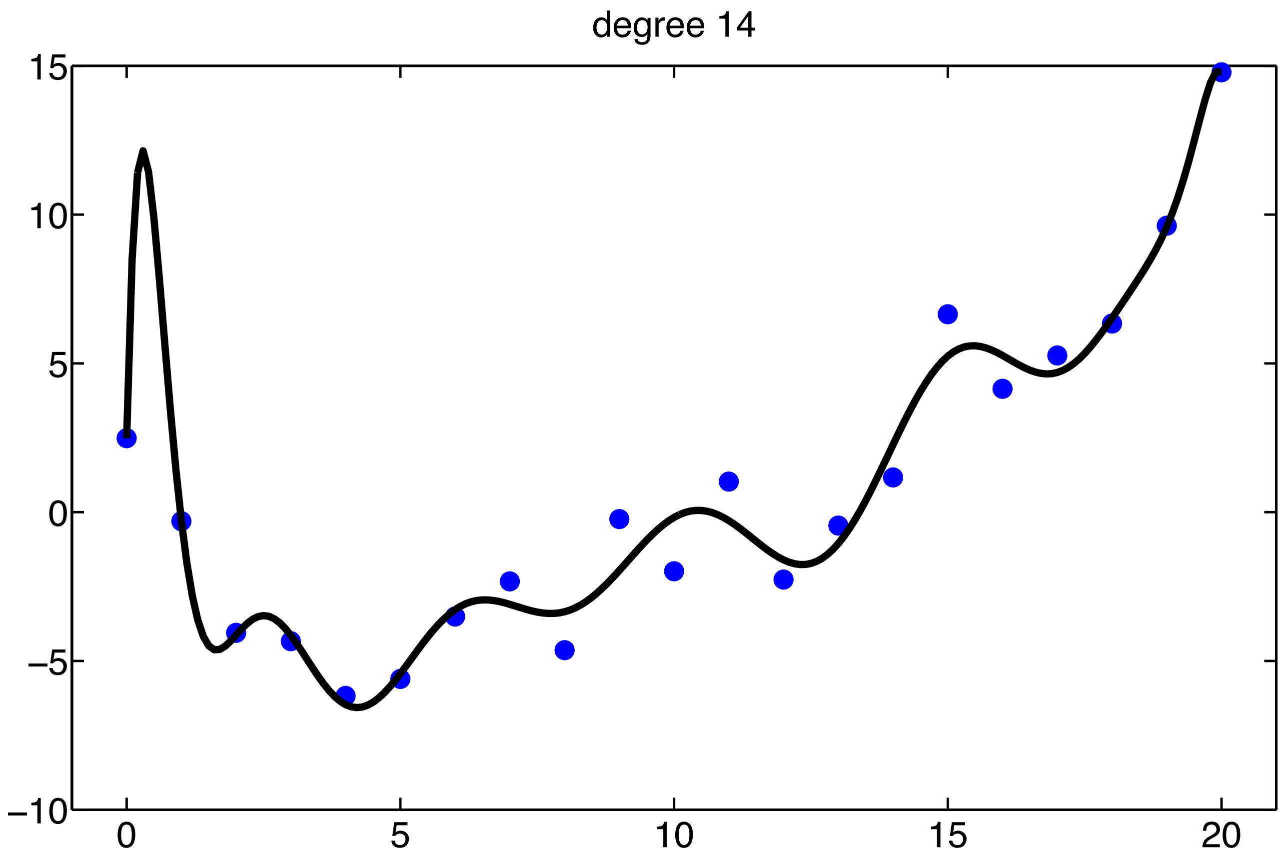
Sum of squares $$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

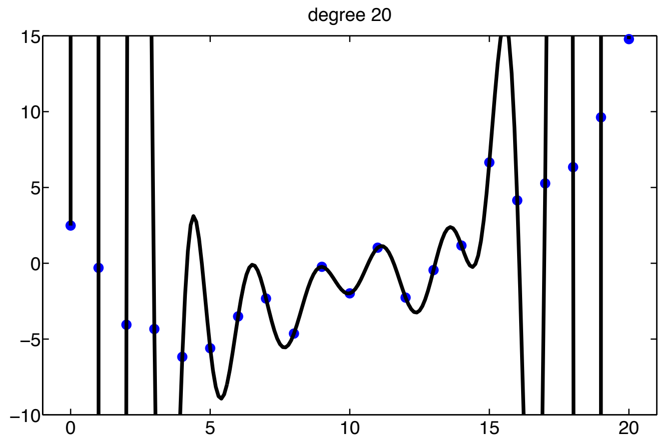


MODEL SELECTION

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



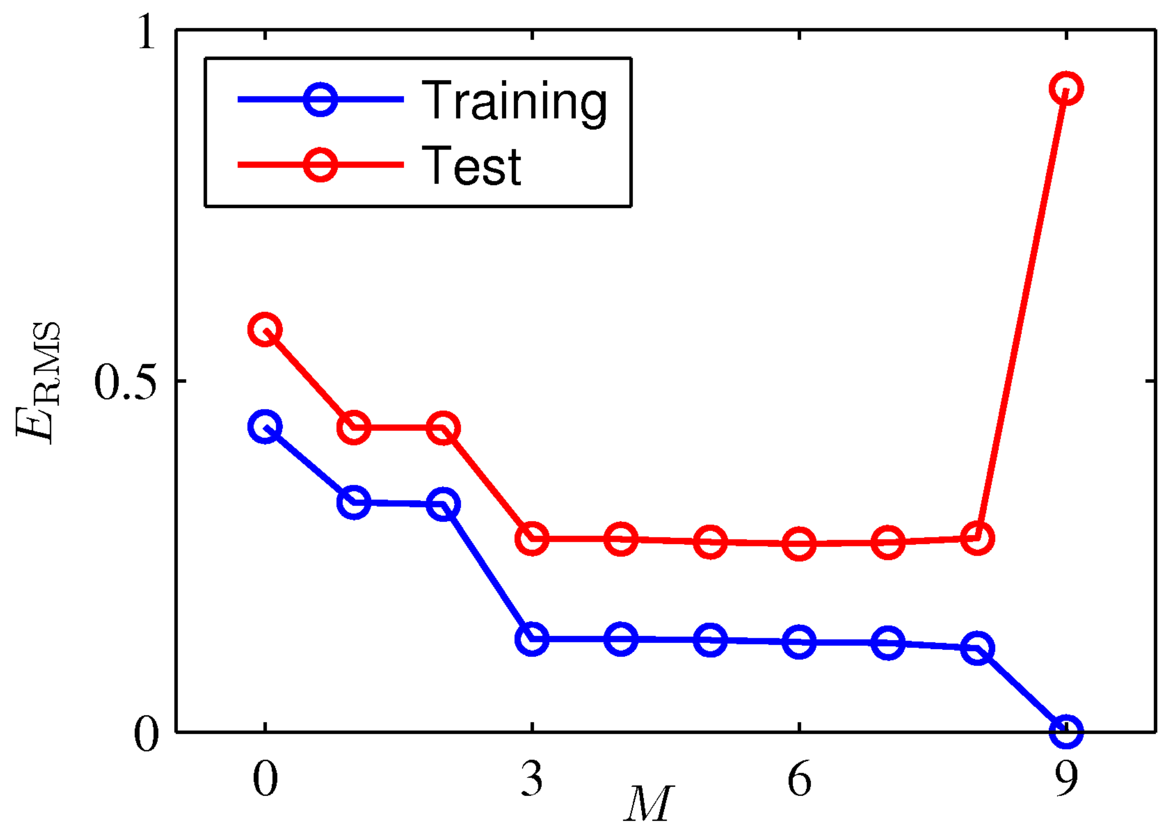




OVERFITTING

COMPARING TEST & TRAINING SETS

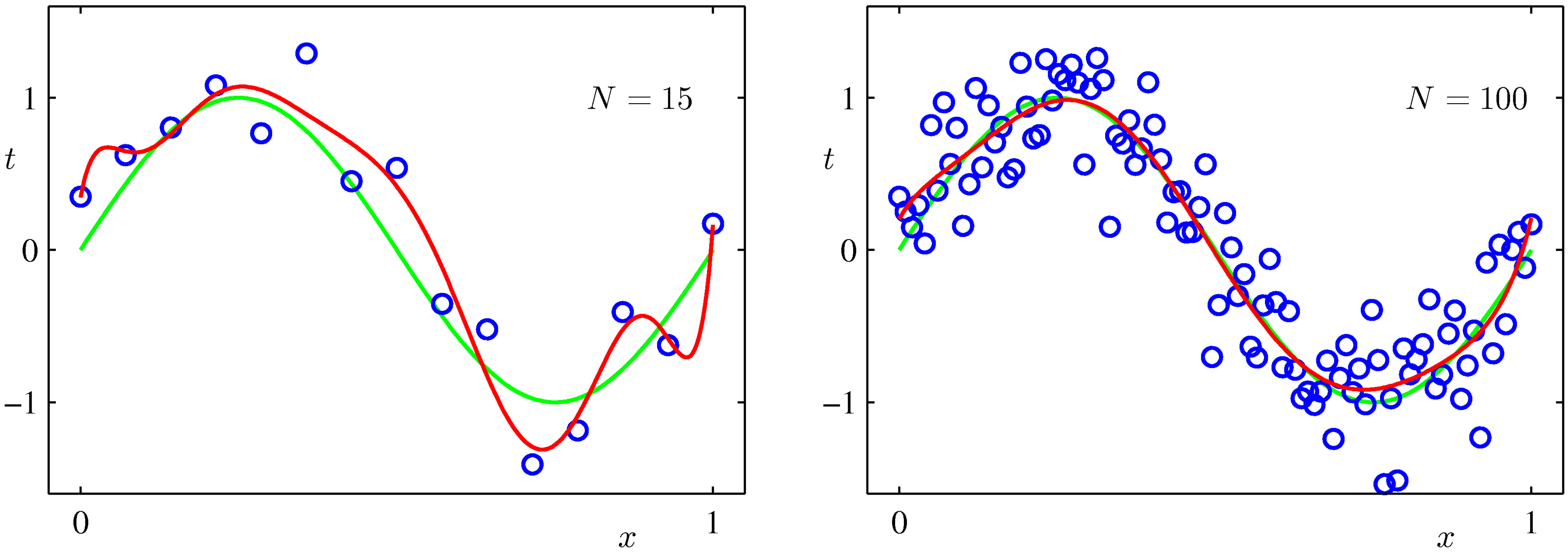
Root-mean-square (RMS) error $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$



SIZE OF COEFFICIENTS

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

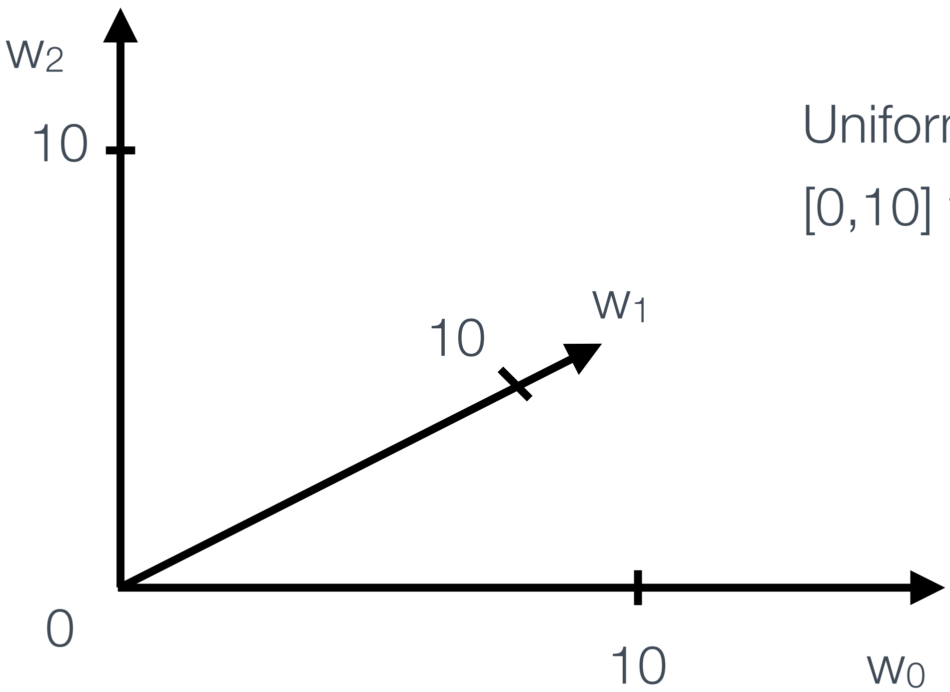
LARGER DATA ALLOWS MORE COMPLEXITY



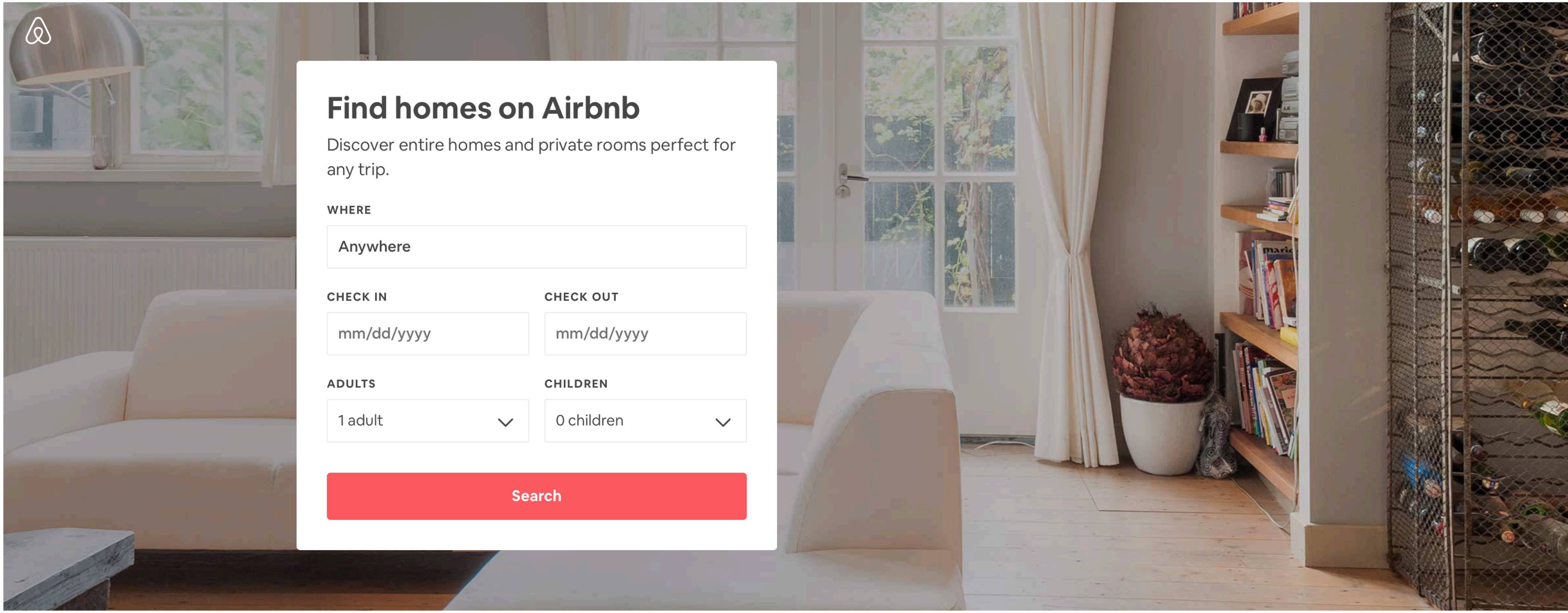
ADDING A THIRD PARAMETER

$$y(x, \mathbf{w}) = w_0$$
$$y(x, \mathbf{w}) = w_0 + w_1 x$$
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2$$

?



Uniform prior over [0,10] for each



Find homes on Airbnb

Discover entire homes and private rooms perfect for any trip.

WHERE

Anywhere

CHECK IN

mm/dd/yyyy

CHECK OUT

mm/dd/yyyy

ADULTS

1 adult

CHILDREN

0 children

Search

What guests are saying about homes in United States

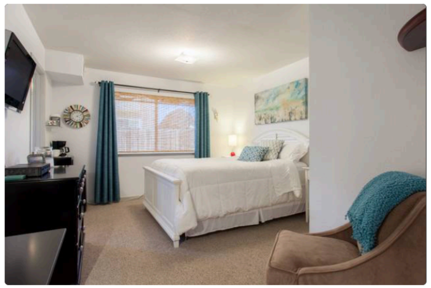
United States homes were rated **4.8 out of 5 stars** with **28,000,000+ reviews**

★★★★★

This place was amazing! Such an Austin experience! Every detail in the airstream was perfect. We wish we could have stayed longer!

★★★★★

We had a very relaxing weekend staying here. Perfect location.



★★★★★

The perfect place to stop on the highway 1 drive. Rosa welcomed us and provided everything that we needed. The space was...

SMART PRICING

Smart Pricing lets you set your prices to automatically match demand, with the goal of attracting bookings. To make sure you're always comfortable with your listing's daily prices, we give you a couple of simple settings to establish the boundaries you're comfortable with:

- The **minimum price** you set is the lowest your price will go when demand for your space is low. This means nightly prices may drop to attract more guests to book, but never below the threshold you set.
- The **maximum price** you set is the highest price your listing can be booked for, even on high demand nights. You can set this as high as $10,000 per night, and it is not publicly displayed. Currently, every listing that uses Smart Pricing must include a maximum price setting.

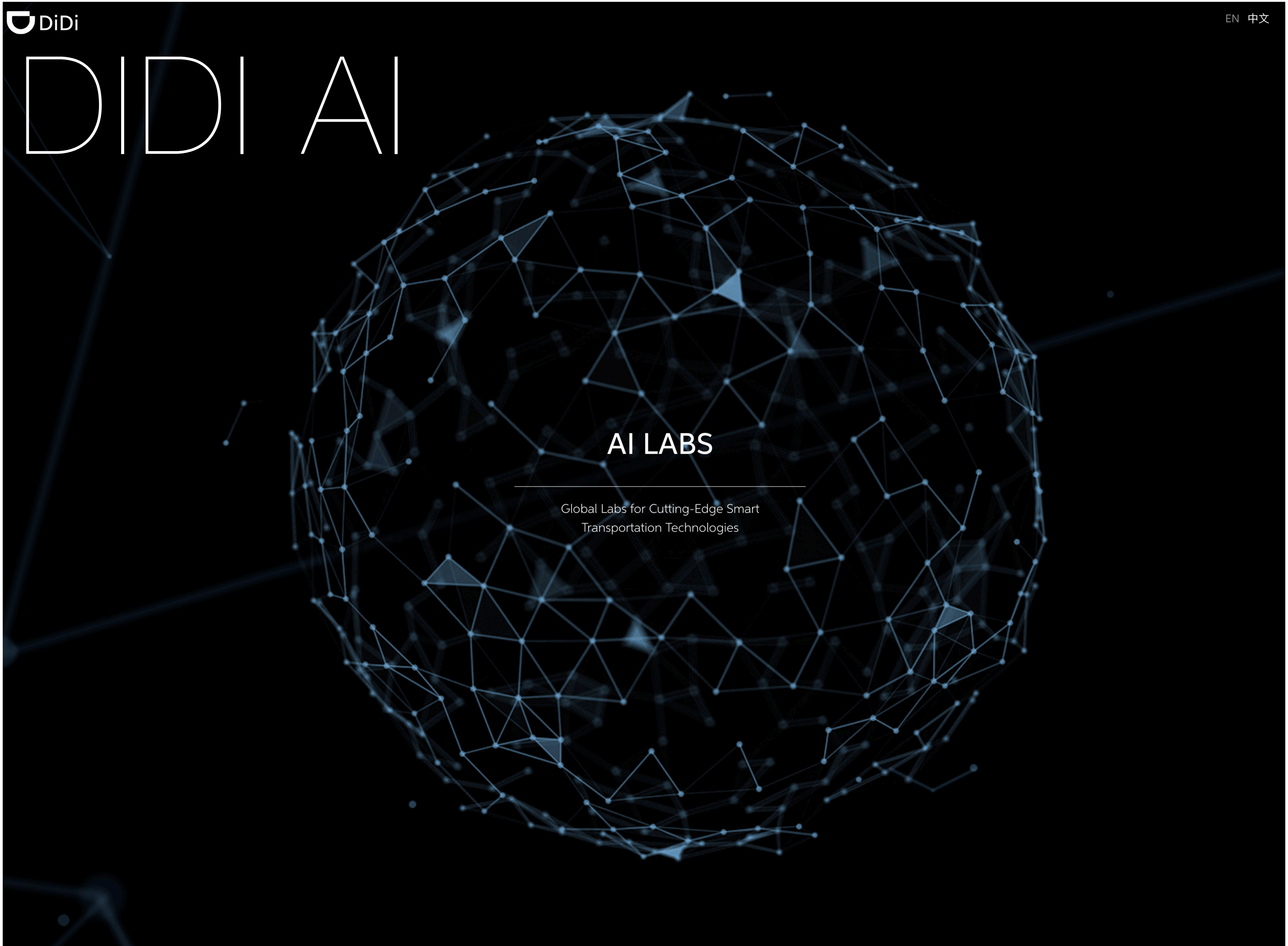
SMART PRICING

When you have Smart Pricing turned on, your pricing suggestions reflect the controls you've set, combined with a lot of data. In fact, Smart Pricing takes into account over 70 different factors that could change your price. These factors, plus your controls, determine the best price for each available night on your calendar, and your price updates to reflect changes in factors like:

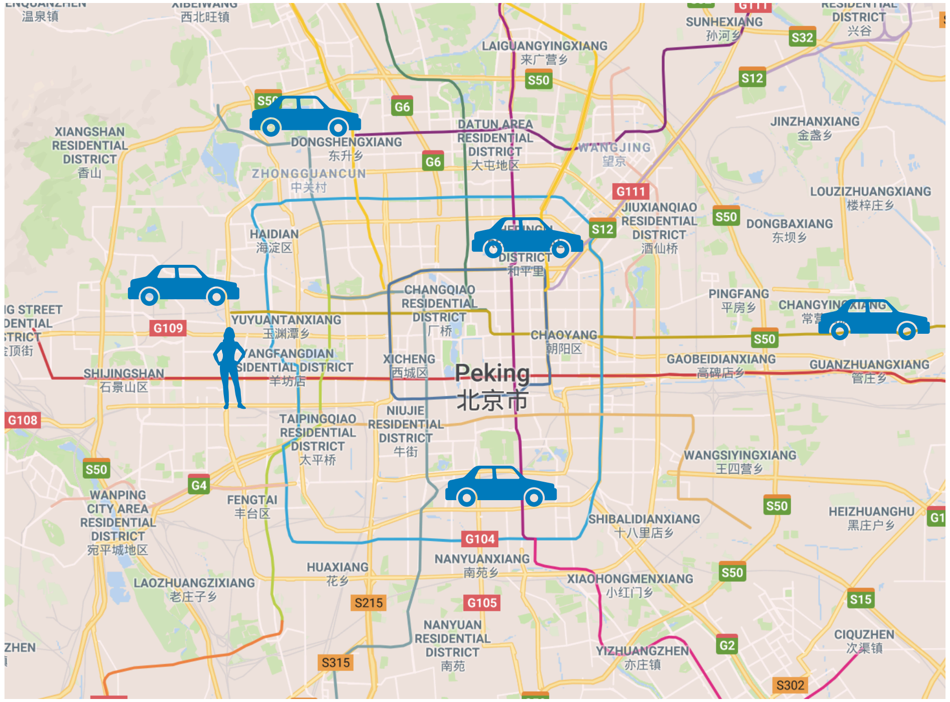
- Lead-time: as a check-in date approaches, your price will update
- Market popularity: if more people are searching for homes in your area, your price will update
- Seasonality: as you move into, or out of high season, your price will update
- Listing popularity: if you get a lot of views and bookings, your price will update
- Listing details: if you add amenities, such as WiFi, your price will update
- Bookings history: as you get bookings, your future prices will be partly based on the prices you got for successful bookings. So, for instance, if you set your price higher than Smart Pricing suggests, and you get a successful booking at that price, the algorithm will update to reflect that.
- Review history: Your prices update as you get more positive reviews from successful stays.

DIDI

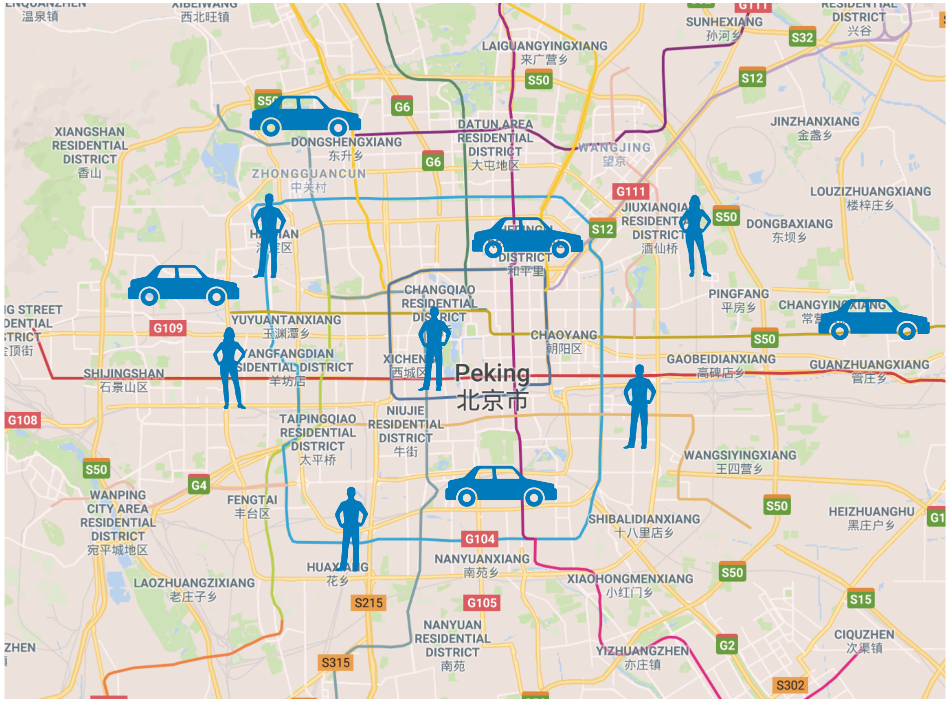

DiDi
EN 中文
ABOUT US
More Than a Journey
A World-Leading Diversified
One-Stop Mobility Platform


DiDi
EN 中文
DIDI AI
AI LABS
Global Labs for Cutting-Edge Smart
Transportation Technologies

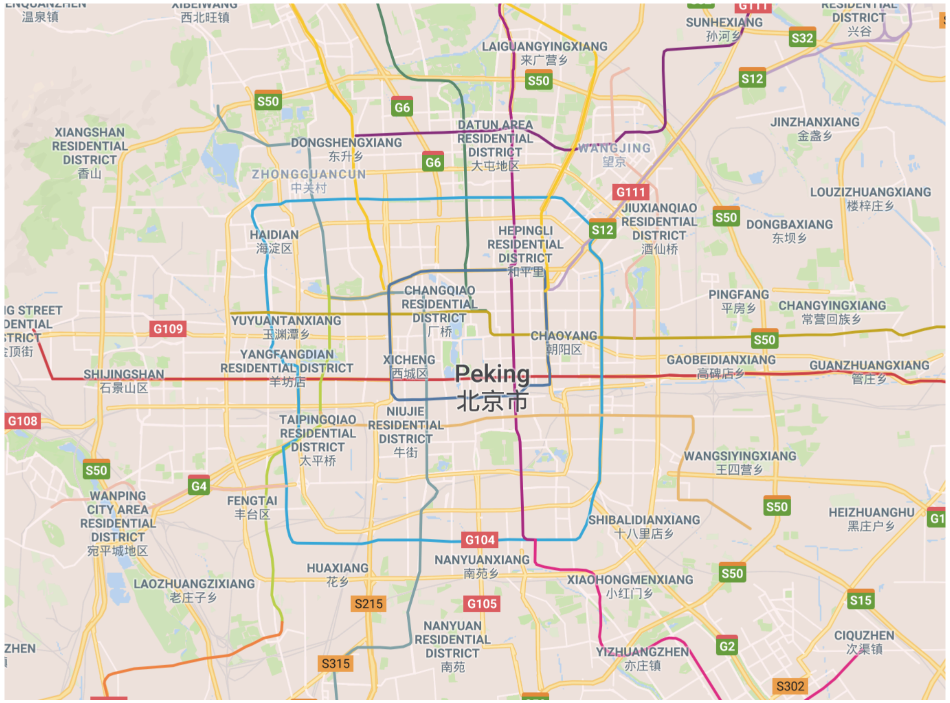
GAUSSIAN PROCESS

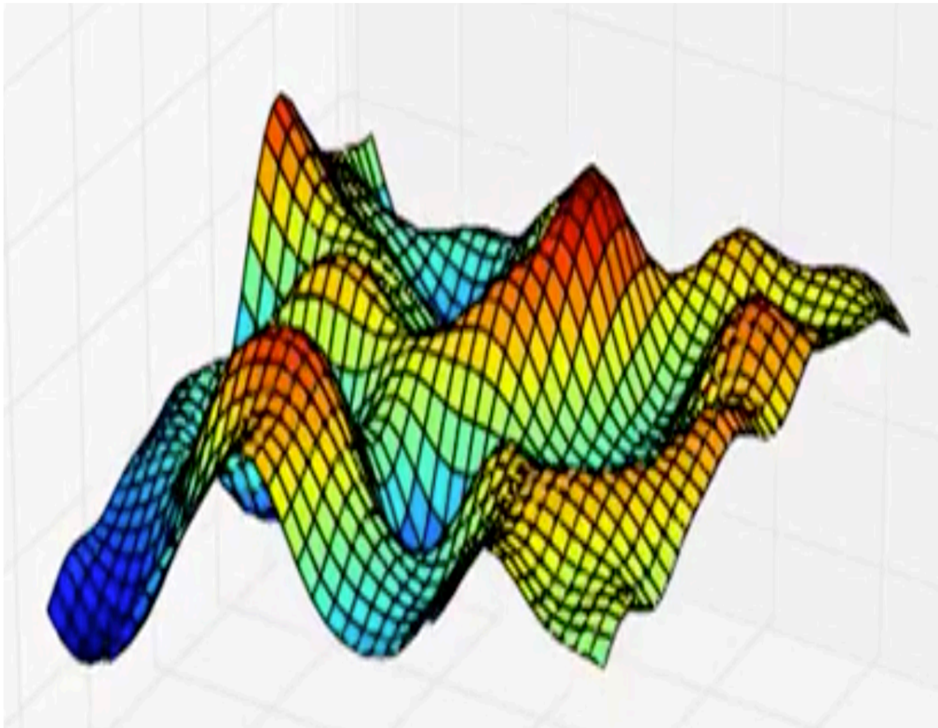


GAUSSIAN PROCESS

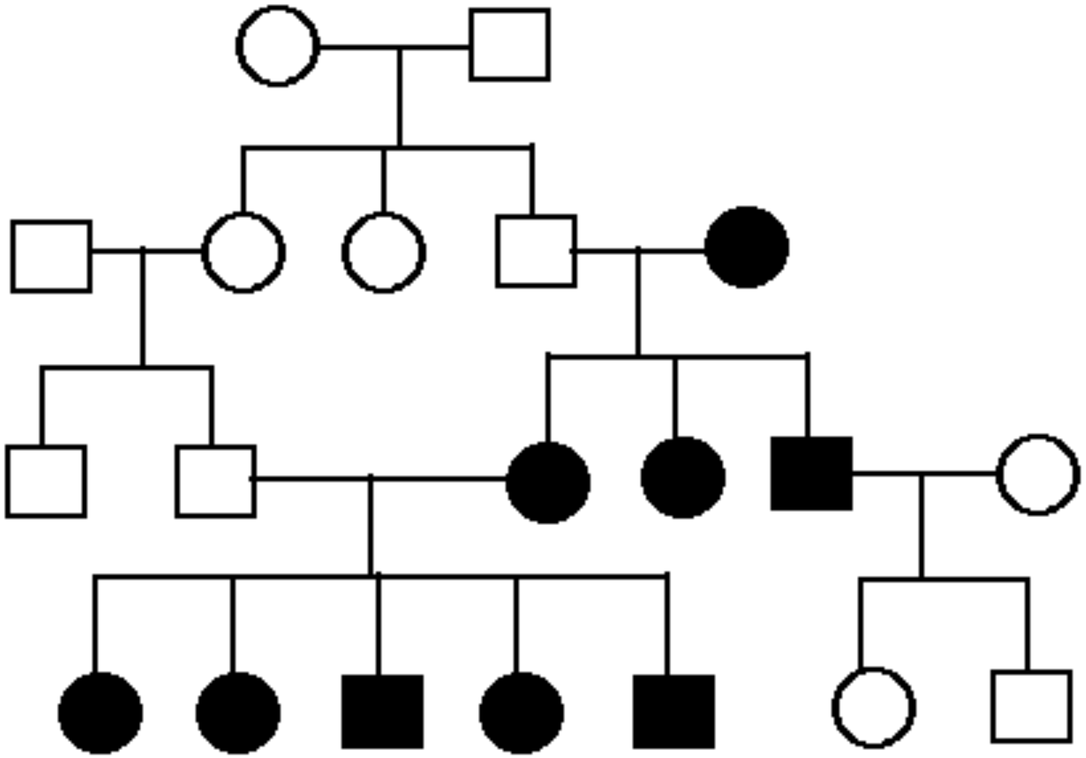


GAUSSIAN PROCESS



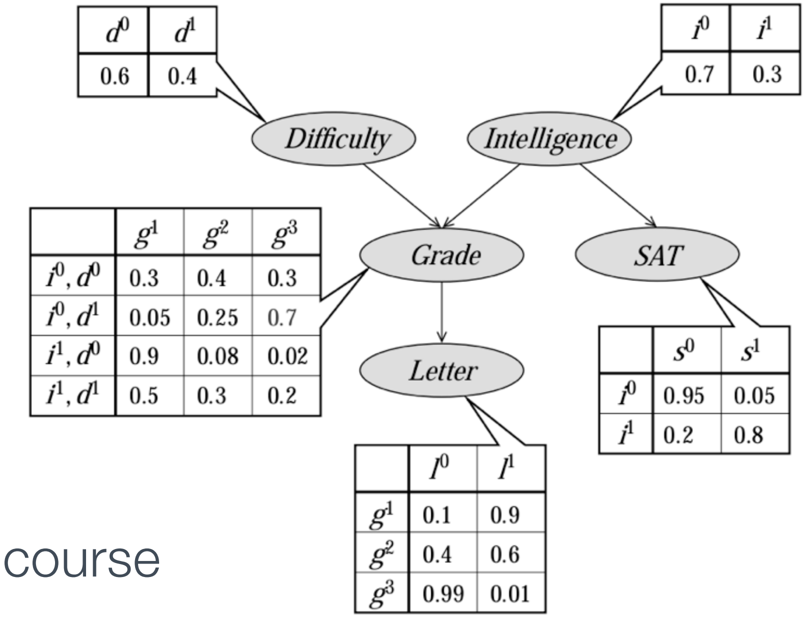


A PEDIGREE



- Three types of problems
  1. Marginalizing
  2. Learning parameters
  3. Learning structure

This course



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

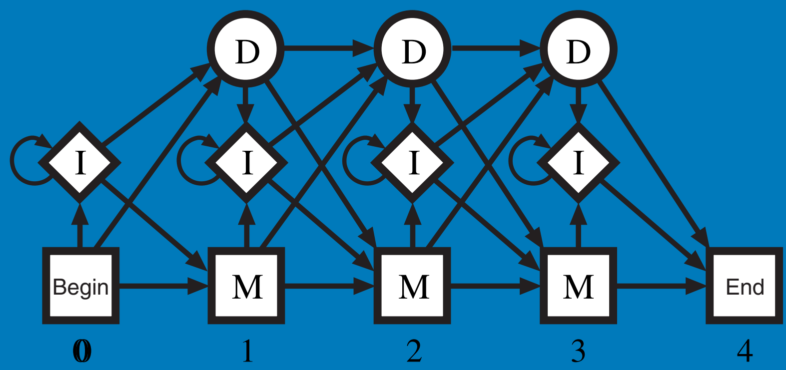
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

DIRECTED GRAPHICAL MODELS

APPLICATIONS OF HMMS

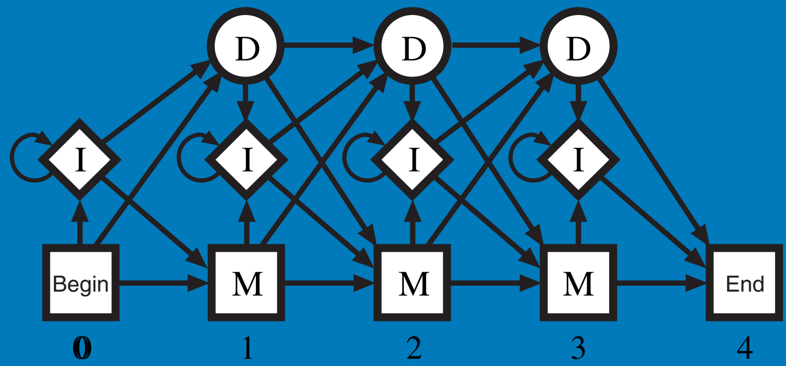
	x	x	.	.	.	x
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3



- Automatic speech recognition
- Part of speech tagging
- Gene finding
- Gene family characterization
- Secondary structure prediction

INFERENCE TYPES

	x	x	.	.	.	x
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3



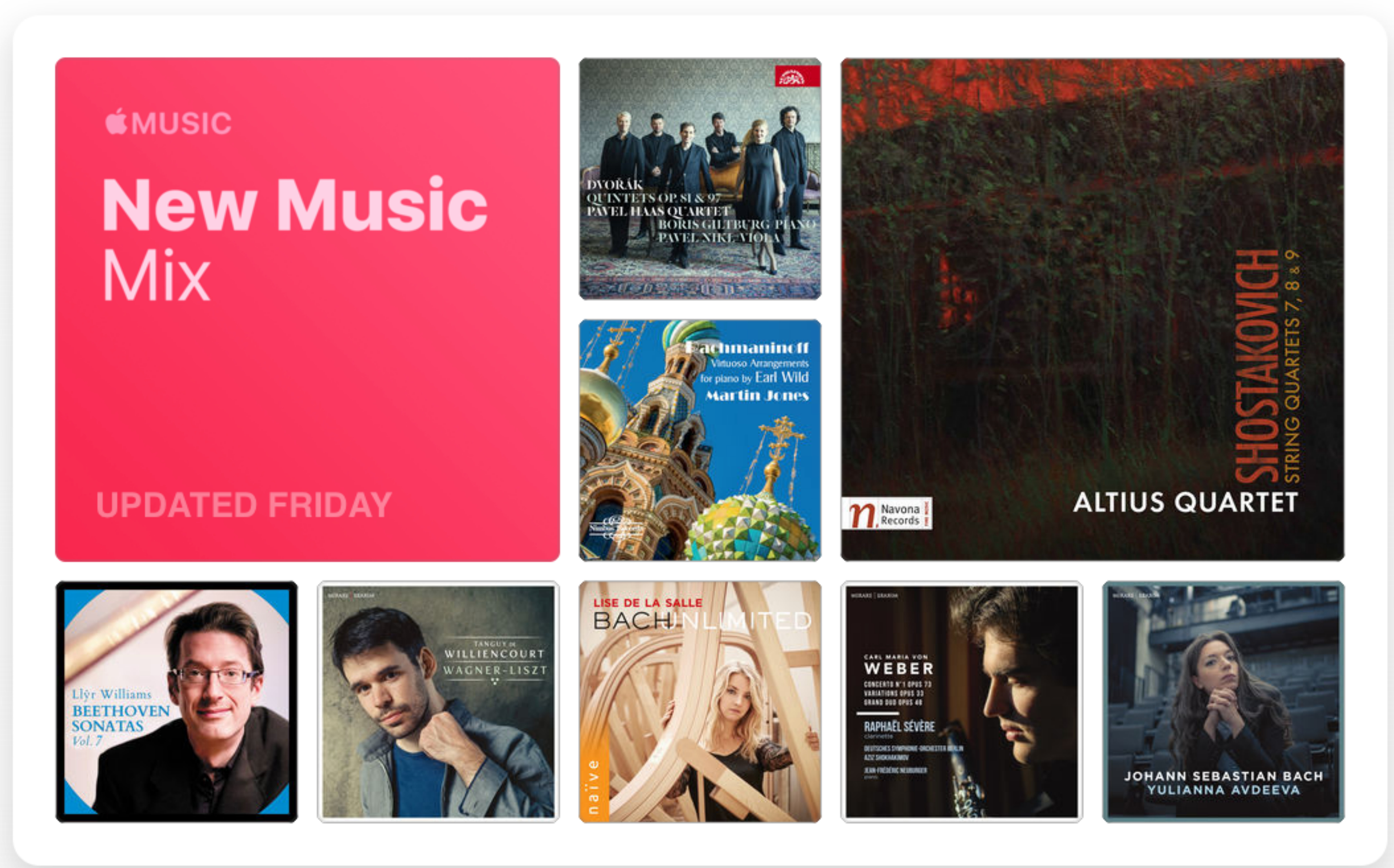
- Probability of data: $p(x_{1:T})$
- Parameters:

  given D & struct.

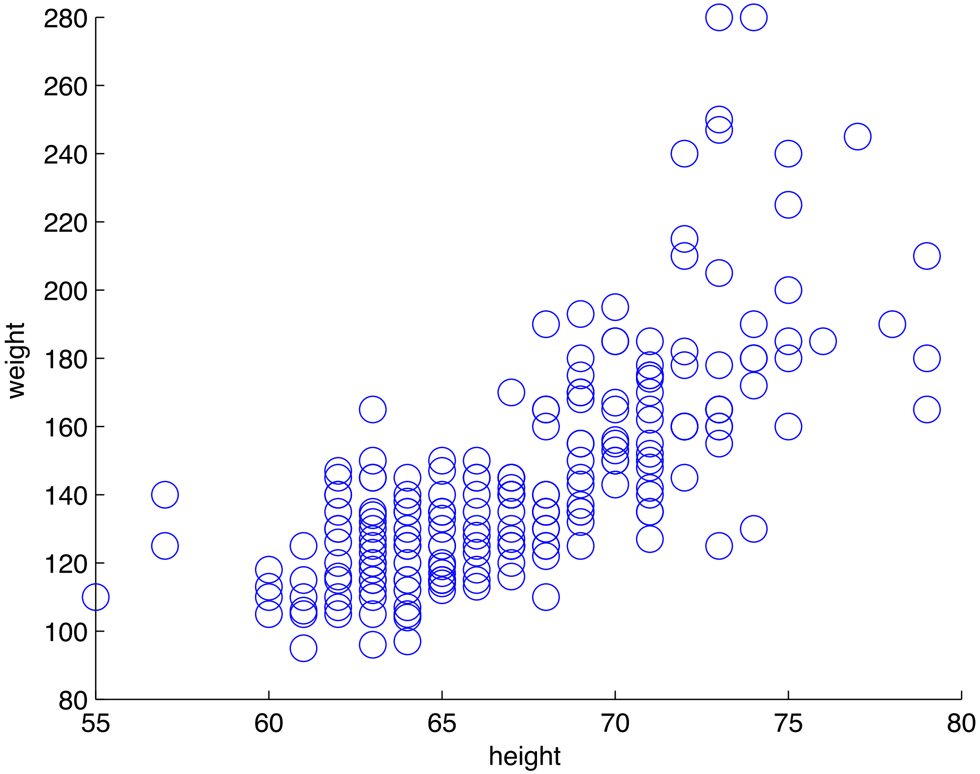
- Structure and param.:

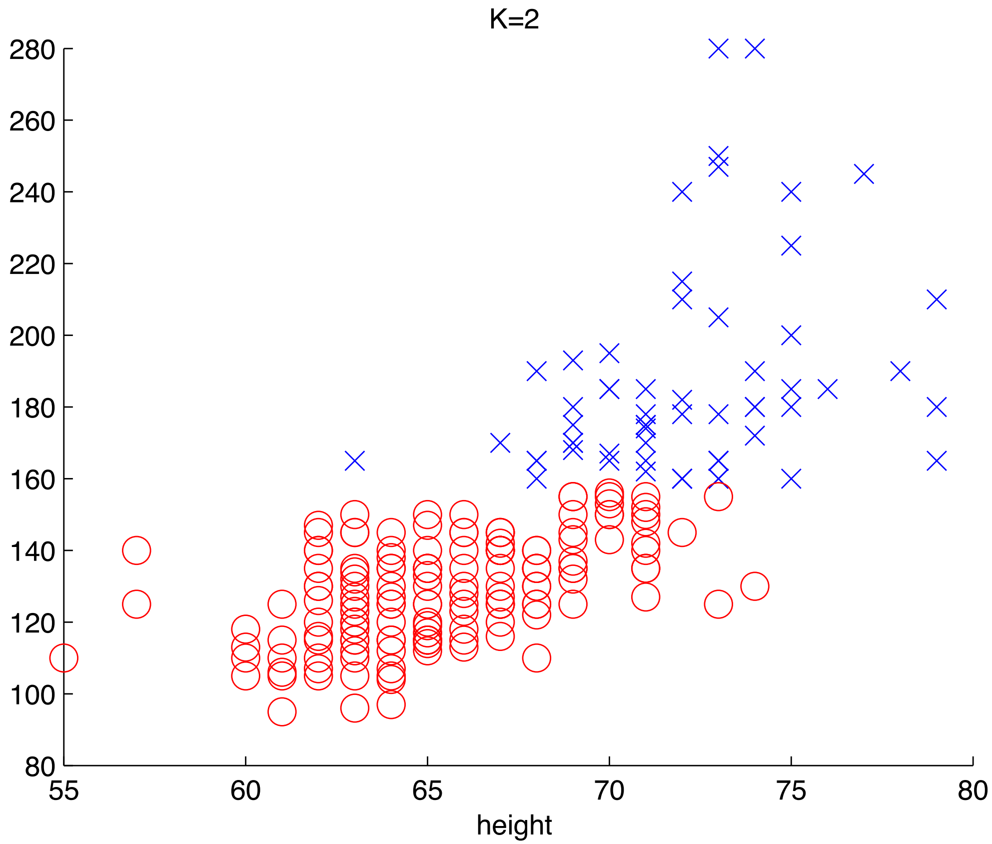
  given D

Using the Expectation Maximisation methodology - EM



COLLABORATIVE FILTERING



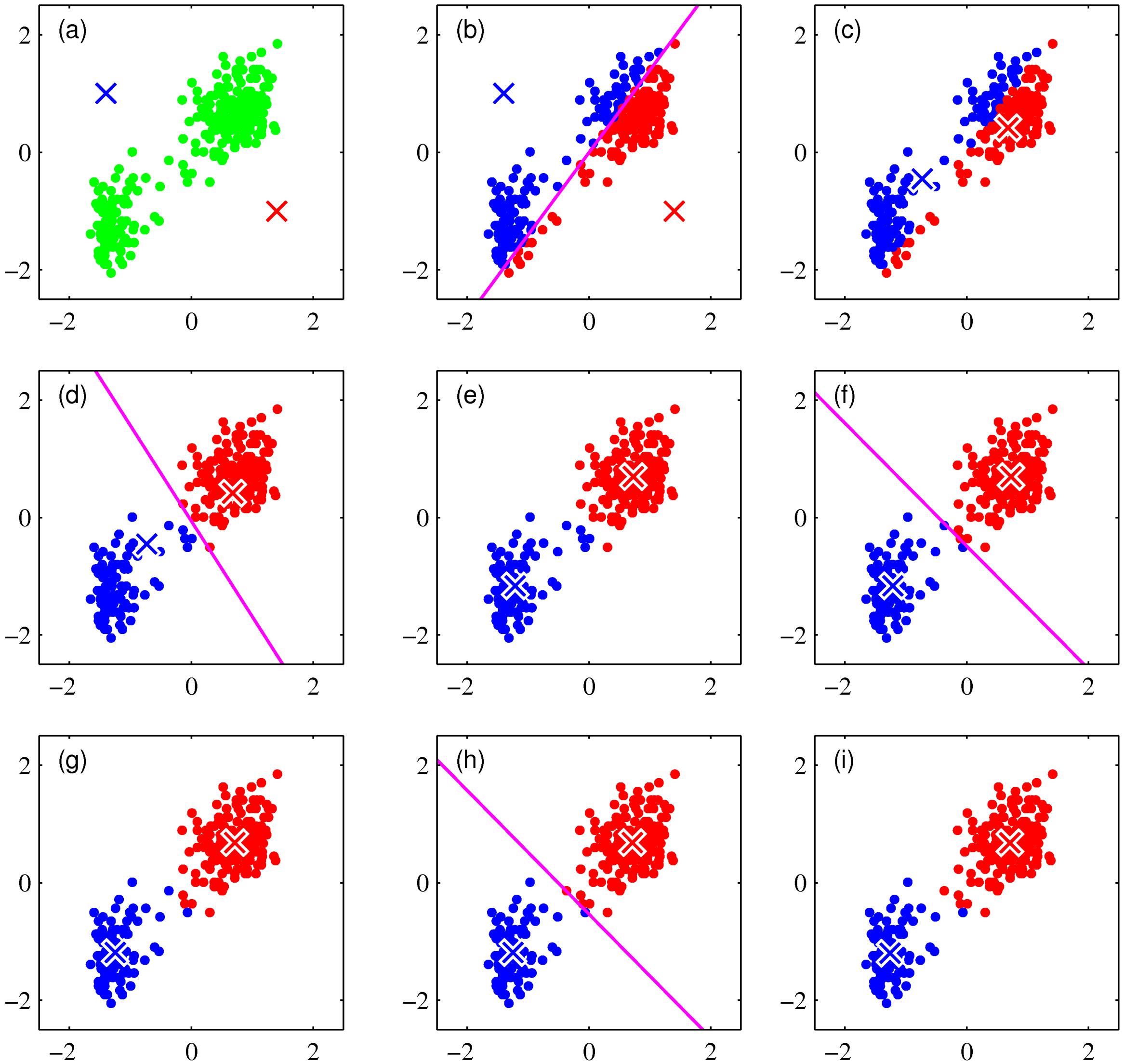


- Each subset should contain similar points

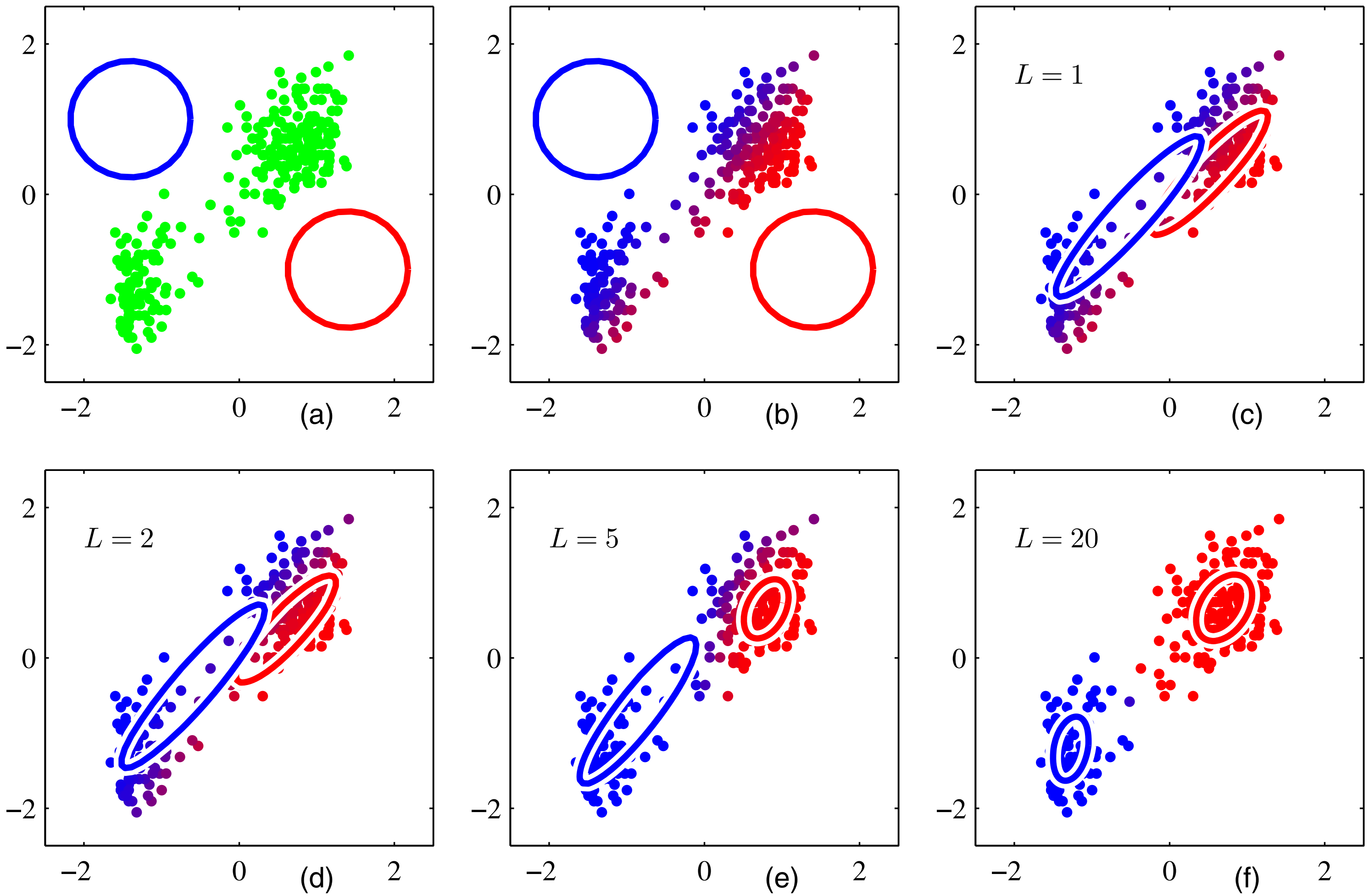
- Pairs of subsets should have dissimilar points.

CLUSTERING

K-MEANS



EM CLUSTERING - CAN BE VIEWED AS SOFT VERSION



VARIATIONAL BAYES

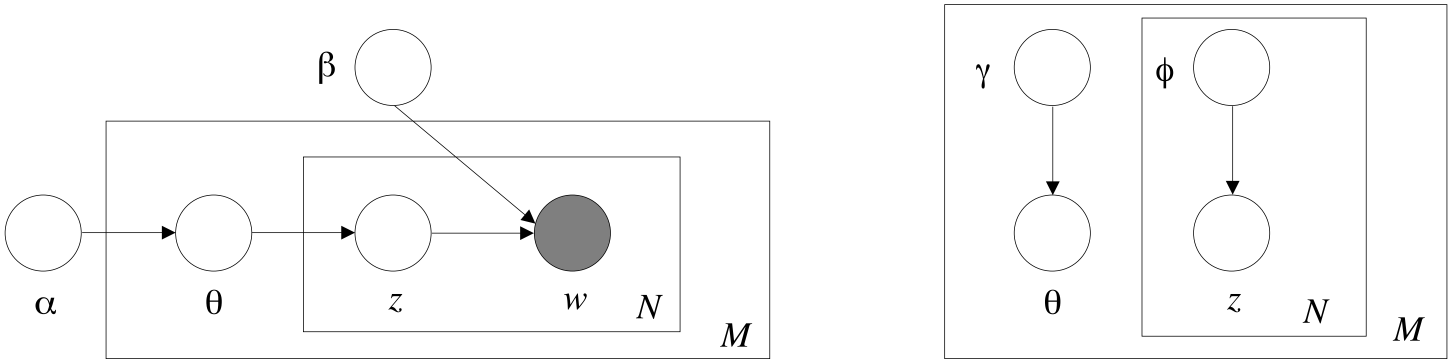


Figure 5: (Left) Graphical model representation of LDA. (Right) Graphical model representation of the variational distribution used to approximate the posterior in LDA.

- Variational Bayes (VB)
    - Technique for approximating a posterior

CONTENT

- Chapter 1, Intro: The entire chapter.
- Chapter 2, Probability Distributions: The entire chapter is background but you need to know a lot of this, in particular 2.3 and 2.4.
- Chapter 3, Linear Models for Regression: The entire chapter, (emphasis on sections 3.1, 3.3 and 3.4).
- Chapter 6, Kernel Methods: 6.1,6.2, and 6.4
- Chapter 8, Graphical Models: 8.1-8.3
- Chapter 9, Mixture Models and EM: 9.1-9.3
- Chapter 10, Approximate Inference: 10.1-10.3.1 except 10.2.2 and 10.2.3.
- Chapter 12, Continuous Latent Variables: focus on sections 12.1-12.3 and 12.4.1
- Chapter 13, Sequential Data: 13.1, 13.2.1, 13.2.2, 13.2.5, 13.2.6

THE END